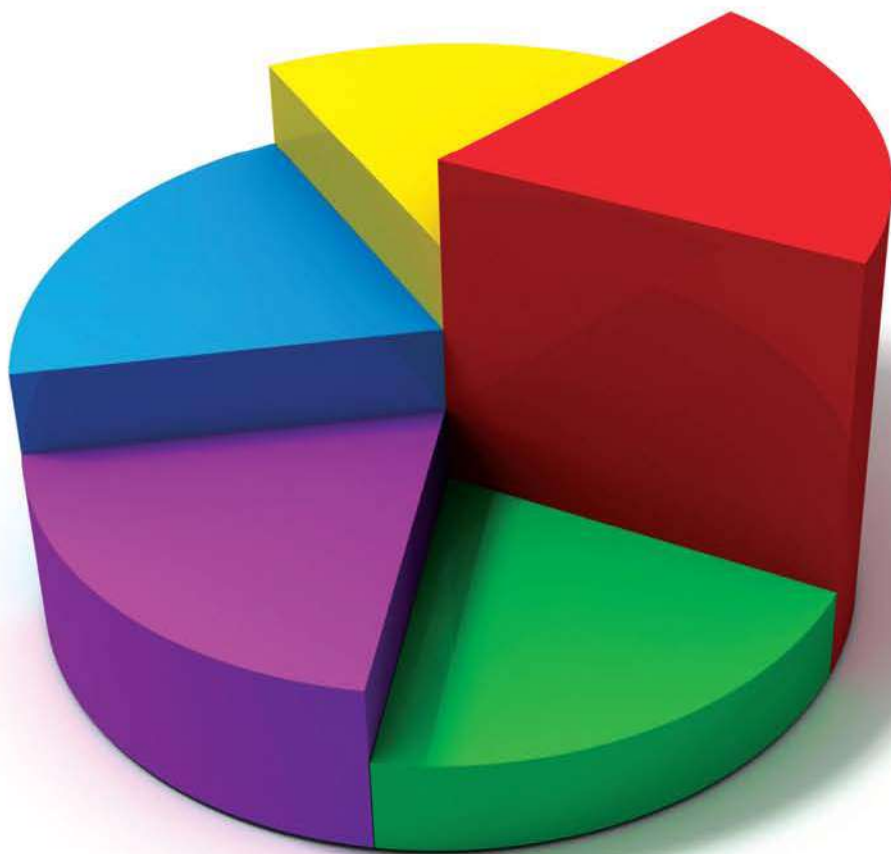


Investigación Comercial

Técnicas e Instrumentos

Rafael Domingo Martínez Carrasco



INVESTIGACIÓN COMERCIAL: **TÉCNICAS E INSTRUMENTOS**

INVESTIGACIÓN COMERCIAL: TÉCNICAS E INSTRUMENTOS

Rafael Domingo Martínez Carrasco



Datos de catalogación bibliográfica:

INVESTIGACIÓN COMERCIAL:

TÉCNICAS E INSTRUMENTOS

Rafael Domingo Martínez Carrasco

EDITORIAL TÉBAR, S.L., Madrid, año 2011

ISBN digital: 978-84-7360-471-0

Materias: 658.8 Organización de mercados. Investigación

Formato: 165 × 240 mm Páginas: 294

www.editorialtebar.com

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con la autorización expresa de Editorial Tébar. La infracción de estos derechos puede ser constitutiva de delito contra la propiedad intelectual (arts. 270 y siguientes del Código Penal).

INVESTIGACIÓN COMERCIAL:

TÉCNICAS E INSTRUMENTOS

© 2011 Editorial Tébar, S.L.

C/ de las Aguas, 4

28005 Madrid (España)

Tel.: 91 550 02 60

Fax: 91 550 02 61

pedidos@editorialtebar.com

www.editorialtebar.com

ISBN digital: 978-84-7360-471-0

Diseño editorial: Carmen Mateos

Diseño de portada: CMYK, S.L.

*A mi pequeño Álex,
bienvenido a casa*

Índice

TEMA 1: INTRODUCCIÓN A LA INVESTIGACIÓN COMERCIAL	13
1. LA INVESTIGACIÓN COMERCIAL	13
2. ETAPAS DE LA INVESTIGACIÓN COMERCIAL	14
3. FUENTES DE INFORMACIÓN.....	16
4. LA ENCUESTA Y LA ENTREVISTA.....	18
TEMA 2: LA INVESTIGACIÓN EXPLORATORIA.....	23
1. TIPOS Y FORMAS DE INVESTIGACIÓN	23
2. LA INVESTIGACIÓN EXPLORATORIA	24
3. VENTAJAS Y DESVENTAJAS DE LA INVESTIGACIÓN EXPLORATORIA	25
4. TÉCNICAS DE INVESTIGACIÓN EXPLORATORIA	26
TEMA 3: LA INVESTIGACIÓN CONCLUYENTE	33
1. LA INVESTIGACIÓN CONCLUYENTE	33
2. TIPOS DE INVESTIGACIÓN CONCLUYENTE.....	34
3. VENTAJAS Y DESVENTAJAS DE LA INVESTIGACIÓN CONCLUYENTE.....	35
4. MÉTODOS DE INVESTIGACIÓN CONCLUYENTE	36
TEMA 4: LA ENCUESTA Y EL CUESTIONARIO	41
1. INTRODUCCIÓN	41
2. LA ENCUESTA	42
3. TIPOS DE ENCUESTA	42
4. FASES EN LA ELABORACIÓN DE UNA ENCUESTA.....	43
5. FORMAS DE REALIZAR UNA ENTREVISTA.....	43
6. EL CUESTIONARIO	45
7. DISEÑO DEL CUESTIONARIO	46
8. TIPOS DE PREGUNTAS.....	48
9. EL CONTROL DE LOS EVALUADORES Y LOS GESTORES	53

TEMA 5: LA TABULACIÓN DE ENCUESTAS	59
1. LAS VARIABLES ESTADÍSTICAS	59
2. TABULACIÓN DE ENCUESTAS	60
TEMA 6: EL MUESTREO.....	67
1. TEORÍA DEL MUESTREO	67
2. TIPOS DE MUESTREO.....	67
3. TAMAÑO DE LA MUESTRA Y ERROR DE MUESTREO.....	70
EJERCICIOS TEMA 6.....	73
TEMA 7: ESTADÍSTICAS DESCRIPTIVAS SIMPLES	75
1. INTRODUCCIÓN A LA ESTADÍSTICA.....	75
2. CLASIFICACIÓN DE LA ESTADÍSTICA	75
3. TIPOS DE PRESENTACIÓN DE DATOS EN ESTADÍSTICAS DE UNA VARIABLE	76
4. DISTRIBUCIÓN DE FRECUENCIAS	78
5. REPRESENTACIONES GRÁFICAS	78
6. PARÁMETROS DESCRIPTIVOS	83
EJERCICIOS TEMA 7.....	92
TEMA 8: LA ESTADÍSTICA BIVARIANTE.....	99
1. TABLAS DE PRESENTACIÓN DE DATOS BIVARIANTES.....	99
2. DIAGRAMA DE DISPERSIÓN	100
3. REGRESIÓN	101
4. AJUSTE POR MÍNIMOS CUADRADOS.....	102
5. VARIANZA RESIDUAL Y COEFICIENTE DE DETERMINACIÓN	193
6. CORRELACIÓN LINEAL SIMPLE	104
EJERCICIOS TEMA 8.....	106
TEMA 9: LAS SERIES TEMPORALES.....	109
1. INTRODUCCIÓN	109
2. ANÁLISIS DE UNA SERIE TEMPORAL.....	109
3. CÁLCULO DE LA TENDENCIA.....	111
4. MEDIAS MÓVILES	113
5. VARIACIÓN ESTACIONAL.....	115
EJERCICIOS TEMA 9.....	119

TEMA 10: LOS NÚMEROS ÍNDICE	123
1. INTRODUCCIÓN	123
2. ÍNDICES SIMPLES.....	123
3. ÍNDICES COMPUESTOS	124
4. UTILIDADES DE LOS NÚMEROS ÍNDICE	125
EJERCICIOS TEMA 10	126
TEMA 11: CONTRASTE DE HIPÓTESIS.....	129
1. DEFINICIÓN DE INFERENCIA ESTADÍSTICA.....	129
2. LAS HIPÓTESIS.....	129
3. ANÁLISIS INFERENCIAL UNIVARIABLE.....	130
4. ANÁLISIS INFERENCIAL BIVARIABLE.....	137
EJERCICIOS TEMA 11	145
TEMA 12: EL INFORME DE INVESTIGACIÓN COMERCIAL.....	149
1. INTRODUCCIÓN	149
2. TIPOS DE INFORME.....	149
3. FORMATO DEL INFORME.....	150
4. ESTRUCTURA DE LOS INFORMES	151
5. PRESENTACIÓN DE DATOS	154
SOLUCIONES A LOS EJERCICIOS.....	155

Tema I

Introducción a la investigación comercial

I. La investigación comercial

Antes de pasar al estudio de la investigación comercial, es necesario resaltar que los conceptos de investigación comercial e investigación de marketing se consideran sinónimos. Sin embargo, no ocurre lo mismo con los conceptos de investigación comercial e investigación de mercados. La diferencia radica en que dentro de la investigación comercial, además de la investigación de mercados, cabe incluir la investigación de publicidad, de distribución, etc. Es pues un concepto más amplio.

La American Marketing Association (A.M.A.) define la investigación comercial como “la obtención, clasificación y análisis de todos los hechos y datos acerca de los problemas relacionados con la transferencia y venta de mercancías y servicios de productos al consumidor”.

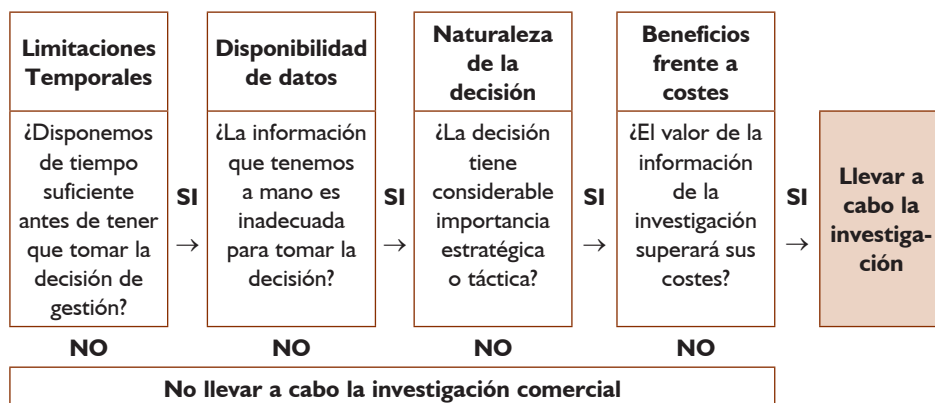
La investigación comercial consiste en la búsqueda y obtención de los datos necesarios para la planificación comercial.

Es conveniente la realización de una investigación comercial cuando se cumplan una serie de condiciones:

- **Limitaciones temporales:** habría que determinar si existe tiempo suficiente para realizar la investigación antes de tomar la decisión para la cual investigamos.
- **Disponibilidad de los datos:** será conveniente realizar la investigación cuando los datos disponibles actualmente no son suficientes para tomar una decisión.
- **Naturaleza de la decisión:** la investigación comercial supone el consumo de recursos (económicos, personales...). Habrá que valorar si la

decisión a tomar con la información obtenida es suficientemente importante como para compensar el empleo de esos recursos.

- **Beneficios frente a costes:** toda investigación comercial tiene un coste económico. La decisión de realizar una investigación comercial ha de estar valorada económicamente para que los beneficios que se obtengan de ella superen sus costes.



2. Etapas de la investigación comercial

Las etapas de la investigación comercial son las siguientes:

1. Definición del problema: la finalidad de esta etapa es que el investigador capte de una forma adecuada lo que se pretende conocer y, de esta forma, elaborar un planteamiento acorde con el tema objeto de la investigación comercial.

2. Análisis de la situación actual: consiste en buscar, dentro de la empresa, toda la información escrita disponible sobre el problema o que pueda aportar datos sobre él.

3. Investigación preliminar: el objeto de esta fase es obtener información de todas las personas de la empresa o cercanas a ella que puedan estar relacionadas con el problema. A estos efectos se realizarán entrevistas de carácter informal a estas personas.

Con la información obtenida en estas tres etapas se estará en disposición de elaborar una hipótesis sobre el problema. Si la causa del problema ya se ha averiguado, la empresa puede considerar innecesario continuar la investiga-

ción comercial. Ahora bien, si desea contrastar los resultados con la opinión del mercado o bien la empresa no tiene una idea clara sobre la causa del problema, se considera conveniente continuar la investigación comercial.

4. Preparación y programación: esta etapa debe ser realizada por personal especializado. Se deben fijar unos objetivos, que se pretenderán obtener con la investigación comercial.

Posteriormente se deben determinar las técnicas a utilizar: encuesta, entrevista en profundidad, reunión en grupo (normalmente, entre 6 y 12 personas), panel de consumidores (muestra fija a la que se investiga periódicamente sobre un tema concreto para estudiar su evolución), estudio ómnibus (es una encuesta realizada para varios clientes al mismo tiempo), técnicas de observación (las personas investigadas no se dan cuenta de ello, con lo que no existen condicionantes) y técnicas de experimentación (reproducen situaciones reales para poder prever resultados, problemas, etc.).

Una vez determinada la técnica a utilizar se debe definir la muestra de la población a la que se va a aplicar. Una muestra es un número de unidades pertenecientes a la población que representen lo más posible las características del universo. Es imprescindible diseñar el tamaño de la muestra y ello dependerá del error que se está dispuesto a tolerar, el presupuesto y el coste.

Posteriormente, se define el cuestionario, en caso de haber elegido como técnica la realización de una encuesta. El cuestionario es el documento en el que se recogen las preguntas que se realizarán durante el desarrollo de la encuesta.

5. Desarrollo de la investigación: se denomina también “trabajo de campo”. Es la fase más importante de la investigación, pues de la veracidad de la información obtenida depende el éxito de la investigación.

La selección y contratación del equipo de entrevistadores es esencial. Generalmente la empresa suele seleccionar los encuestadores y formarlos adecuadamente, dándoles instrucciones no sólo sobre el trabajo en general, sino sobre el estudio concreto que se esté realizando.

Después de la recogida de información es necesario realizar un control que permita descubrir los errores básicos que puedan invalidar total o parcialmente la información recogida.

6. Estudio analítico e interpretación de datos: el objeto de esta fase es la de convertir la información obtenida en información utilizable. Para ello se pueden utilizar métodos estadísticos de síntesis.

7. Informe: se suelen presentar dos informes: el informe de conclusiones, cuya finalidad es permitir una rápida lectura de los resultados más importantes del estudio, y el informe técnico, que debe ser lo más completo y exhaustivo posible, comenzando con una introducción sobre los motivos que originaron el estudio. A continuación se presentarán los objetivos que se perseguían, las fuentes y los datos utilizados, las técnicas aplicadas, el diseño de la muestra, su validación, los resultados obtenidos, los problemas surgidos, cómo se han resuelto, y un ejemplar de los cuestionarios que se hayan utilizado.

3. Fuentes de información

Las fuentes de información de la investigación comercial pueden clasificarse en dos grandes grupos:

- **Fuentes primarias:** son las informaciones obtenidas directamente por la empresa, utilizando las herramientas de investigación propias de esta disciplina. La obtención de esta información suele ser más costosa, ya que la propia empresa es la que realiza las labores de obtención de la misma.
- **Fuentes secundarias:** es la información que ya existe acerca del objeto a estudiar, información que podemos encontrar en la propia empresa (documentos, estudios anteriores, experiencia...) o fuera de ella (revistas, prensa, internet...).

Comenzaremos analizando las fuentes secundarias, para abordar después las primarias. En este tema de introducción ofrecemos un análisis general de las fuentes primarias, que serán objeto de más detalle en los temas posteriores.

3.1. Fuentes secundarias

Ventajas

Mayor disponibilidad, rapidez y economía que los datos de encuestas, tests y observación.

Desventajas

La investigación se diseñó a la medida de otra investigación y puede no adecuarse a lo que buscamos.

Fuentes

- a) Fuentes estadísticas**
- b) Organismos oficiales y empresariales**
- c) Organizaciones de consumidores y medios de comunicación**
- d) Empresas de estudios de mercado e información comercial**

3.2. Fuentes primarias

a) Experimento

Se trata de realizar una prueba a pequeña escala, controlando las condiciones para manipular una o varias variables. Cambiando distintas variables se comprueba si existe relación entre el precio y las ventas, las ventas y la ubicación de los productos, etc.

Puede realizarse tanto en laboratorio (por ejemplo, pruebas ciegas, pruebas cobaya, etc.) como en el exterior (pruebas de campo en comercios, en la calle, etc).

Ventajas

Al ser una prueba a pequeña escala permite reproducir las condiciones reales antes de ser implantadas a nivel de todo el mercado o usuarios.

Inconvenientes

- Posibles sesgos del investigador principal.
- Aparición de variables no controladas, no esperadas.
- Representatividad de los miembros escogidos.

b) Observación

Se trata de observar el comportamiento del mercado o de los consumidores ante cualquier hecho.

Entre otras, la observación se podría realizar acerca de:

- Observación del comportamiento humano: comportamiento no verbal frente al uso de un producto-servicio.

- Cantidad de personas que frecuentan un establecimiento.
- Tipo de compra que realizan.
- Registro detallado de eventos que suceden según la hora del día y el día de la semana.
- N° de peatones que circulan frente a un establecimiento comercial.
- N° de vehículos que transitan por una calle o zona comercial.
- Comentarios que realizan los consumidores frente al uso de un producto-servicio.

Ventajas

Podemos obtener información sobre el comportamiento de las personas y objetos.

Inconvenientes

Las actitudes, motivaciones y expectativas no son observables y puede tener el sesgo interpretativo de la persona que observa.

c) Encuesta

Se trata de pedir información a una muestra representativa de personas, denominados encuestados, utilizando preguntas escritas. Los cuestionarios o entrevistas recopilan datos cara a cara, por teléfono, por correo o través de medios de comunicación.

Es el mejor modo de averiguar lo que el consumidor piensa. Dada la imposibilidad de tiempo y económica de entrevistar a todos los posibles miembros de la población, encuestamos solamente a una parte representativa. Nace el error de muestreo que disminuye a medida que aumenta la muestra.

4. La encuesta y la entrevista

4.1. Entrevista personal

Recopila información a través del contacto cara a cara con los individuos.

Ventajas

- Posible interacción para que no haya malinterpretaciones.
- Sondeo de respuestas complejas.

- Longitud de la entrevista.
- Integridad del cuestionario: reducción de las respuestas en blanco.
- Material auxiliar y ayuda visual: presentación de muestras, bocetos y otras ayudas.
- Alta participación: presencia del encuestador.

Inconvenientes

- Influencia del entrevistador: tono de voz, apariencia...
- Falta de anonimato del encuestado.
- Alto coste.

4.2. Entrevista telefónica

Recopila la información a través del contacto telefónico con los encuestados.

Ventajas

- Velocidad de recopilación de datos.
- Coste: se eliminan tiempos y gastos de desplazamiento.
- Cooperación: se coopera más en entrevistas telefónicas.

Inconvenientes

- Teléfonos móviles: aumento de hogares sin teléfono fijo y falta de ficheros con sus números de teléfono.
- Cuestionarios cortos: no sobrepasan los 5-6 minutos.
- Ausencia de ayudas visuales.

4.3. Encuesta postal

Cuestionario autoadministrado que se remite por correo al encuestado. Actualmente está la modalidad del correo electrónico.

Ventajas

- Flexibilidad geográfica: posibilidad de alcanzar una muestra dispersa.
- Costes: bajo coste.

- Comodidad del encuestado: según el tiempo del encuestado.
- Anonimato del encuestado: respuestas confidenciales.
- Ausencia de entrevistador: en casos de respuestas no deseadas socialmente o delicadas.
- Longitud: permite que sea un cuestionario largo.

Inconvenientes

- Tasa de respuesta baja: en torno al 10-15%.
- Tiempo de respuesta: se alarga en el tiempo.

4.4. Encuesta por Internet

Cuestionario autoadministrado emplazado en un sitio web.

Ventajas

- Velocidad y rentabilidad: gran audiencia mundial en tiempo record.
- Participación del encuestado: el usuario navega a ese sitio web para participar.
- Anonimato del encuestado: permiten preguntas de tipo confidencial y delicado.

Inconvenientes

- Muestra representativa: no suele representar al total de la población especialmente si no disponen de conexión a Internet o no están familiarizados con el ordenador y las nuevas tecnologías.

4.5. Dinámicas de grupo

Es una reunión poco estructurada que transcurre con un pequeño grupo de personas. No hay un cuestionario estructurado sino un guión con los temas a tratar.

Ventajas

- Temas a tratar: se puede hablar casi de cualquier cosa, con múltiples perspectivas.
- Matización de respuestas: se pueden conocer los sentimientos de la gente, sus frustraciones y opiniones.

- **Rapidez:** en poco tiempo puede obtenerse respuestas y fácil ejecución.

Inconvenientes

- **Moderador:** hace falta un moderador experimentado si se quiere obtener un resultado óptimo.
- **Coste:** el alquiler de la sala, el incentivo a los participantes, el moderador, la captación de los invitados, etc. tiene como consecuencia un alto coste.
- **Representatividad:** no permite que sus respuestas sean representativas del total de la población.

4.6. Entrevista en profundidad

Es una entrevista extensa con poca estructuración en la que el entrevistador pregunta muchas cuestiones y sondea respuestas exhaustivas. Es similar a la que podría hacer un psicólogo a un paciente.

Ventajas

- **Temas a tratar:** se puede hablar casi de cualquier cosa, con múltiples perspectivas.
- **Matización de respuestas:** se pueden conocer los sentimientos de la gente, sus frustraciones y opiniones.

Inconvenientes

- **Entrevistador cualificado:** hace falta un entrevistador cualificado para realizar este tipo de entrevista.
- **Coste:** el desplazamiento del técnico especialista al lugar de desplazamiento y el coste de su tiempo.
- **Representatividad:** no permite que sus respuestas sean representativas del total de la población.

Tema 2

La investigación exploratoria

1. Tipos y formas de investigación

Las investigaciones pueden ser clasificadas:

a) Según la naturaleza de los objetivos en cuanto al nivel de conocimiento que se desea alcanzar.

- **La investigación exploratoria:** es considerada como el primer acercamiento científico a un problema. Se utiliza cuando éste aún no ha sido abordado o no ha sido suficientemente estudiado y las condiciones existentes no son aún determinantes.
- **La investigación descriptiva:** se efectúa cuando se desea describir, en todos sus componentes principales, una realidad.
- **La investigación correlacional:** es aquel tipo de estudio que persigue medir el grado de relación existente entre dos o más conceptos o variables.
- **La investigación explicativa o causal:** es aquella que busca una relación causal; no sólo persigue describir o acercarse a un problema, sino que intenta encontrar las causas del mismo.

Las tres últimas pueden englobarse dentro de la llamada investigación concluyente, que pretende suministrar información que ayude a evaluar y seleccionar las líneas de acción.

b) Según el tiempo en que se efectúan:

- **Investigaciones sincrónicas:** son aquellas que estudian fenómenos que se dan en un corto período.
- **Investigaciones diacrónicas:** Son aquellas que estudian fenómenos en un período largo con el objeto de verificar los cambios que se pueden producir.

c) Según la naturaleza de la información que se recoge para responder al problema de investigación:

- **Investigación cuantitativa:** es aquella que utiliza predominantemente información de tipo numérico.
- **Investigación cualitativa:** es aquella que persigue describir sucesos complejos en su medio natural, con información preferentemente cualitativa.

2. La investigación exploratoria

La investigación exploratoria tiene como principal objetivo ofrecer una comprensión inicial del problema a tratar. Esta fase permite definir el problema de una manera más precisa, a través de la cual se plantearán diferentes hipótesis que orienten el desarrollo de posteriores investigaciones.

La información obtenida a través de esta investigación ayudará a dar un mayor entendimiento del problema facilitando el desarrollo de una investigación posterior. Sin embargo, debido a que la investigación exploratoria no es estructurada, estos datos no se tomarán como concluyentes. En esta etapa, para obtener información, el investigador comenzará con el análisis de datos secundarios para posteriormente realizar un compendio de los datos primarios.

Este tipo de investigación puede hacerse por diversos motivos:

- a) Dirigidos a la formulación más precisa de un problema de investigación, dado que se carece de información suficiente y de conocimientos previos del objeto de estudio. En este caso, la exploración permitirá obtener nuevos datos y elementos que pueden conducir a formular con mayor precisión las preguntas de investigación.
- b) Conducentes al planteamiento de una hipótesis: cuando se desconoce el objeto de estudio resulta difícil formular hipótesis acerca del mismo. La función de la investigación exploratoria es descubrir las bases y recabar información que permita, como resultado del estudio, la formulación de unas hipótesis. Las investigaciones exploratorias son útiles por cuanto sirven para familiarizar al investigador con un objeto que hasta el momento le era totalmente desconocido. Sirve como base para la posterior realización de una investigación descriptiva, puede crear en otros investigadores el interés por el estudio de un nuevo tema o problema y

puede ayudar a precisar un problema o a concluir con la formulación de una hipótesis.

La investigación exploratoria no intenta dar una explicación respecto del problema, sino sólo recoger e identificar antecedentes generales, números y cuantificaciones, temas y tópicos respecto del problema investigado, sugerencias de aspectos relacionados que deberían examinarse en profundidad en futuras investigaciones. Su objetivo es documentar ciertas experiencias, examinar temas o problemas poco estudiados o que no han sido abordados antes. Por lo general, investigan tendencias, identifican relaciones potenciales entre variables y establecen el “tono” de investigaciones posteriores más rigurosas.

La necesidad de una investigación exploratoria surge cuando se necesita:

- Obtener información sobre la posibilidad de llevar a cabo una investigación más completa sobre un contexto particular de la vida real. Este tipo de estudio pretende generar datos e hipótesis que constituyen la materia prima para investigaciones más precisas.
- Investigar comportamientos que se consideran cruciales.
- Identificar conceptos o variables concretas.
- Establecer prioridades para investigaciones futuras.
- Sugerir afirmaciones (postulados) verificables.

No hay un campo metodológico desarrollado para las investigaciones exploratorias. En general, este tipo de investigaciones se caracterizan por la gran flexibilidad que ofrecen en su metodología, ya que ésta puede ser cuantitativa o cualitativa, según sean las necesidades que lleva a realizar una investigación de este tipo.

3. Ventajas y desventajas de la investigación exploratoria

3.1. Ventajas

- La investigación exploratoria es, posiblemente, el mejor camino para hacer el planteamiento de las hipótesis.
- Ayuda a determinar tendencias e identificar relaciones potenciales entre variables.

- Se caracterizan por ser más flexibles en su metodología en comparación con los estudios descriptivos o explicativos, y son más amplios y dispersos.
- Sirven para aumentar el grado de familiaridad con fenómenos relativamente desconocidos.
- Utilidad inmediata para recoger información para proyectos.

3.2. Desventajas

- Generalmente, la investigación exploratoria no es el mejor mecanismo para tomar decisiones.
- Los estudios exploratorios en pocas ocasiones constituyen un fin en sí mismos.
- Implican un mayor “riesgo” y requieren gran paciencia, serenidad y receptividad por parte del investigador.
- Como investigación de campo es poco relevante y no se la considera seriamente como investigación científica o académica.

4. Técnicas de investigación exploratoria

Como se ha mencionado, no existe una metodología determinada para la investigación exploratoria. A pesar de ello, sin ánimo de ser exhaustivos, se presentan a continuación algunas posibles técnicas de este tipo de investigación.

4.1. Técnicas proyectivas

Consisten en la utilización de vagos, ambiguos e informales estímulos o situaciones, para conocer las características del mundo del individuo o su comportamiento en él. Dentro de los test proyectivos encontramos:

1. El test de libre asociación de palabras

Consiste en presentar personalmente a las personas entrevistadas una lista de palabras diversas relacionadas con el tema de estudio, para que responda rápidamente con cualquier palabra que se le ocurra.

Las contestaciones se clasifican según los siguientes criterios:

- a) Frecuencia de las respuestas comunes: si una palabra aparece en muchas ocasiones, denota la existencia de una cierta actitud de los individuos hacia la misma.
- b) Vacilaciones: como se pide una respuesta rápida, si ésta se produce después de los 3 segundos se considera que ha habido una vacilación. Esta vacilación demuestra una situación emocional del individuo que le impide responder rápidamente.
- c) Falta de contestación: si el entrevistado no puede responder es que la palabra no le transmite realmente ningún mensaje.

Esta técnica se suele utilizar para dar nombre a nuevos productos y para pequeños mensajes publicitarios.

2. El test de frases incompletas

Se presentan frases incompletas para que los encuestados las terminen. Se suelen entrevistar alrededor de 200 personas y proporciona información sobre las actitudes, sentimientos y creencias de las personas (ejemplos: “lo importante de los refrescos es...”, “en verano las bebidas refrescantes...”, “cuando tengo mucha sed...”, etc.).

3. El test de apercepción temática (T.A.T.)

Consiste en presentar 20 láminas con dibujos que representan difusamente determinadas personas o situaciones. El sujeto entrevistado debe contar una historia sobre cada una de las láminas.

4. El test de Rosenzweig

Consta de 24 dibujos que reflejan situaciones de la vida diaria, relacionados con la materia en estudio. El entrevistado debe escribir lo que a su parecer están pensando cada uno de los personajes que aparecen en los dibujos. Suele presentarse en forma de cómic.

4.2. Entrevistas en profundidad

Las entrevistas en profundidad son una forma no estructurada e indirecta de obtener información, pero a diferencia de las sesiones de grupo, las entrevistas se realizan con una sola persona. Este tipo de técnica en la investigación puede tener una duración desde 30 minutos hasta más de una hora, dependiendo del tema y la dinámica de la entrevista. Para ello se requiere la habilidad de un entrevistador que provoque un ambiente de confianza con el

entrevistado a fin de que hable con libertad de sus actitudes, creencias, sentimientos y emociones. Dentro de una entrevista en profundidad es posible combinar técnicas proyectivas a fin de profundizar en algún tema o de obtener respuestas que muchas veces el entrevistado no está dispuesto de forma racional y espontánea a proporcionar.

Puede ser una entrevista centrada sobre el problema (también llamada “entrevista semiestructurada”), consistente en una entrevista planificada por el director de la investigación (que no es el entrevistador), debiendo el entrevistador ceñirse a las indicaciones recibidas de éste, mientras que el entrevistado debe responder simplemente a las preguntas formuladas. Suele durar entre 30 y 90 minutos. El entrevistador no necesita poseer conocimientos especiales.

También puede ser una entrevista centrada sobre la persona (también llamada “entrevista no directa”), en la que la investigación se centra en las experiencias vividas por el individuo, con total libertad en los comentarios. Implica una gran preparación psicológica del entrevistador, ya que debe controlar la entrevista sin que lo parezca. Suele durar entre 2 y 3 horas.

Suele ser habitual realizar entre 40 y 100 entrevistas, ya que la información adicional que se obtiene con un número mayor suele ser insignificante y aumenta notablemente los costes. Estas entrevistas suelen ser grabadas.

4.3. Técnicas en grupo

Son un conjunto de métodos basados en reuniones de grupos de personas que, a través de un proceso de comunicación dinámico entre sus miembros, permite obtener una solución o información sobre un determinado problema. Las técnicas de grupo existentes son muy numerosas, aunque las más habituales son: Phillips 66, mesa redonda con o sin interrogador, el *role-playing*, estudio de casos y el *brainstorming*.

Según los especialistas, el tamaño ideal del grupo está comprendido entre 6 y 10 personas. Suelen ser suficientes entre 4 y 6 reuniones, ya que con ellas no se pretende obtener datos extrapolables al total de la población, sino comprender la importancia que las personas conceden a diversos aspectos objeto de estudio. Estas reuniones suelen ser grabadas en video, e incluso, son analizadas por circuito cerrado, de forma que el interesado en la misma puede dar órdenes al director de la reunión (si lo hay), a través de un micrófono que sólo él puede oír, para incorporar una nueva cuestión a analizar.

4.4. La observación cualitativa

La observación tiene como propósito obtener una descripción completa de los casos y formular una mejor comprensión de las variables, como, los precios de la competencia y la actividad de publicidad, o la amplitud de las líneas de clientes que esperan en la cajas de un supermercado, para ayudar a identificar los problemas y oportunidades.

Una observación sistemática puede, además, ser útil complemento para otros métodos. Durante una entrevista personal, por ejemplo, el entrevistador tiene la oportunidad de anotar el tipo, la condición, el tamaño de la residencia, la raza del entrevistado y el tipo de vecindario. Rara vez esta fuente de datos es adecuadamente usada en las encuestas.

Los tipos de observación podemos clasificarlos en los siguientes:

- **Observación directa:** este método es usado a menudo para obtener indicios del comportamiento del consumidor. Existen dos tipos de observación directa:
 - Estructurada, con una forma detallada de registro preparada anticipadamente.
 - No estructurada: el observador puede ser enviado a incorporarse con los clientes en la tienda y a buscar actividades que indiquen problemas de servicio. Ésta es una tarea demasiado subjetiva porque el observador debe seleccionar pocos aspectos para anotar y registrar gran cantidad de detalles.
- **Observación diseñada:** estos métodos pueden considerarse como pruebas proyectivas del comportamiento; es decir, la respuesta de la gente colocada en una situación diseñada revelará algunos aspectos de sus creencias fundamentales, actitudes y motivos. Permite analizar pruebas como variaciones en el espacio de los estantes, en los sabores de los productos y en las ubicaciones de anuncios.
- **Medidas de rastreo físico:** este enfoque implica el registro del “residuo” natural del comportamiento. Estas medidas son rara vez usadas porque requieren de una gran cantidad de ingenio, y generalmente producen una medición poco exacta. Sin embargo, cuando funcionan pueden ser muy útiles. Por ejemplo, el consumo de alcohol en una comunidad puede ser estimado a partir del número de botellas vacías en la basura. O un negocio de reparación de automóviles podría selec-

cionar las estaciones de radio para impulsar su publicidad observando las más populares en las radios de los automóviles que son entregados para reparación.

- **Dispositivos para el registro del comportamiento:** diversos dispositivos han sido diseñados para superar deficiencias particulares de los observadores humanos. El ejemplo más obvio es el contador de tráfico, el cual opera continuamente sin cansarse, y como consecuencia es más barato y probablemente más exacto que los humanos. Por las mismas razones, así como por la no obstrucción, las cámaras pueden ser usadas en lugar de los humanos. Existen ciertas situaciones en las que la observación humana es imposible y necesita un instrumento de registro. Así, existen dispositivos disponibles para medir los cambios en la tasa de transpiración como una guía para la respuesta emocional a los estímulos, y los cambios en el tamaño de las pupilas de los ojos de los sujetos, las cuales se presume que indican el grado de interés en el estímulo que está siendo observado. Estos pueden ser utilizados tan sólo en ambientes de laboratorio y frecuentemente presentan resultados ambiguos.
- **La observación participante:** en este tipo de observación el investigador observa y participa, plena o parcialmente, en las actividades que definen los fenómenos estudiados. Por ejemplo, para conocer la forma de reaccionar de un grupo determinado, el observador puede integrarse en él e interactuar como componente del mismo. La investigación participante es un método de investigación útil cuando es difícil realizar una observación discreta, cuando el fenómeno es más fácil de comprender participando en él o cuando no se puede hacer de otra manera.
- **El análisis del contenido:** los artículos de periódicos y revistas, los programas de televisión o de radio, la publicidad y otros medios de comunicación masivos ofrecen muchas veces un material de interés para la observación de marketing. Este método consiste en seleccionar una muestra de comunicaciones (por ejemplo, de publicidad impresa) y describir esas comunicaciones por medio de categorías que conciernen la forma (como el número de palabras, la presencia de fotografías) y el contenido (particularmente los argumentos empleados, los temas discutidos). Las categorías están definidas en función de los objetivos per-

seguidos por el investigador, y la codificación de la comunicación se hace con la ayuda de una plantilla.

4.5. La pseudocompra o *mystery shopper*

La pseudocompra, también llamada cliente oculto o cliente fantasma, es una técnica en la que el investigador se presenta en una empresa como un cliente potencial y se comporta como un comprador normal, aunque en realidad está actuando de forma premeditada.

El objetivo de la pseudocompra es analizar cómo reacciona normalmente el vendedor de una empresa. El informe se suele realizar a la salida del establecimiento ya que es el momento en que la información está más fresca y en ese informe se refleja:

- La actitud del vendedor.
- Los argumentos de venta que ha utilizado el vendedor.
- Las marcas ofrecidas al cliente.
- Las soluciones dadas a los problemas planteados por el falso comprador.
- El aspecto interior y exterior del local, así como las características personales del vendedor y su apariencia.
- El movimiento de clientes en ese lugar.

Un rasgo importante de la pseudocompra es que no existe cuestionario ni guión, sino que el entrevistador tiene que estar altamente cualificado para saber qué es lo realmente importante.

Tema 3

La investigación concluyente

1. La investigación concluyente

La investigación concluyente suministra información que ayuda a evaluar y seleccionar las líneas de acción dentro de una investigación comercial. El diseño de la investigación se caracteriza por procedimientos formales, al contrario que ocurriría en la investigación exploratoria. Esta formalidad comprende necesidades definidas de objetivos e información relacionados con la investigación.

Con frecuencia, se redacta un cuestionario detallado, junto con un plan formal de muestreo. Debe ser evidente que la información que se va a recolectar esté relacionada con las alternativas en evaluación.

El objetivo de esta fase de investigación es probar las hipótesis planteadas en la investigación exploratoria y analizar las relaciones entre las variables que influyen en el comportamiento de los consumidores. La investigación concluyente, a diferencia de la investigación exploratoria, es formal y estructurada, por lo que los descubrimientos son considerados como la información de entrada para el proceso de toma de decisiones. En esta investigación el trabajo se realiza con una muestra grande, que sea representativa de la población a estudiar, y los datos son analizados cuantitativamente.

Para el desarrollo de esta etapa se usará fundamentalmente la investigación descriptiva, a través de la cual se pretende describir las particularidades del mercado, dándole al investigador la posibilidad de identificar las interrelaciones existentes entre las diferentes variables que le afectan, y determinar el comportamiento de los consumidores respecto a un producto o servicio, lo que permitirá hacer una segmentación del mercado teniendo en cuenta las cualidades y características propias de la población.

Los posibles enfoques de investigación incluyen encuestas, experimentos, observaciones cuantitativas y los paneles de consumidores. Para realizar el

estudio se tomará una muestra representativa de la población objetivo, la cual será evaluada a través de un estudio concreto.

2. Tipos de investigación concluyente

2.1. Investigación causal

Busca las relaciones de causa y efecto que existen entre las variables que conforman un problema específico. Una herramienta que se utiliza frecuentemente es la regresión múltiple, que consiste en buscar y medir las relaciones que existen cuando una serie de variables se presentan en conjunto.

Es necesario determinar qué variables son la causa (variables independientes) y qué variables son el efecto (variables dependientes) de un fenómeno. Se trata de determinar la naturaleza de las relaciones entre las variables causales y el efecto que debe pronosticarse.

El método principal de la investigación causal es la experimentación; ésta tiene dos tipos de validación:

- Validación interna: medida de la precisión de un experimento. Mide si la manipulación de las variables independientes, o tratamientos, provoca en realidad los efectos en la(s) variable(s) dependiente(s).
- Validación externa: determinación de si las relaciones de causa y efecto descubiertas en el experimento pueden generalizarse.

2.2. Investigación descriptiva

Se emplea en estudios cuyo propósito es describir las características de una situación o un mercado en particular: ¿quién?, ¿qué?, ¿cuándo?, ¿dónde?, perfil de usuario. No está dirigida a comprobar explicaciones, ni hacer predicciones, ni probar hipótesis, aunque los resultados pueden servir de base para formular hipótesis o para hacer predicciones. Los métodos más utilizados para adelantar investigaciones descriptivas son los estudios estadísticos.

Sirve para proporcionar información sobre la forma en que suceden los fenómenos. Por ejemplo, el investigador busca información acerca de las características de los usuarios de un producto específico, o el grado en que un bien o servicio varía con el tiempo, con el ingreso y con las características generales de los compradores. Tiene como propósito lo siguiente:

- Describir el perfil de grupos objetivo (consumidores, vendedores, organizaciones o áreas de mercado).
- Estimar el porcentaje de unidades que presentan cierto comportamiento en una población específica.
- Determinar cómo se perciben las características del producto.
- Determinar el grado de asociación de variables de mercado.
- Hacer predicciones específicas.

3. Ventajas y desventajas de la investigación concluyente

3.1. Ventajas

- En muchos casos, ofrece la posibilidad de tomar los datos de la investigación sobre datos ya elaborados dentro de la empresa y fuera de ella (fuentes secundarias), lo cual supone un cierto ahorro de costes.
- Suministra información que ayuda al gerente a seleccionar y evaluar una línea de acción.
- La investigación concluyente contiene diversidad de herramientas (encuestas, experimentos, observaciones y paneles) que, si son bien utilizadas, aportan un elevado nivel de precisión.
- Describe fenómenos de marketing y su frecuencia de ocurrencia.
- Determina el grado de asociación de variables de marketing, analizando cómo influye cada variable dentro del entorno.
- Las muestras utilizadas son grandes, lo que supone que son significativas en términos estadísticos.

3.2. Desventajas

- Si se plantea mal el problema puede traer graves consecuencias a la hora de extraer la información.
- Algunos estudios son muy costosos y requieren mucho tiempo.
- Existe desconfianza por parte de muchas personas acerca de las afirmaciones estadísticas a las que se puede llegar, fruto de sabidas manipulaciones que históricamente se han realizado.

- Son procesos más estructurados y formales (inflexibilidad), lo cual puede impedir, en algunos casos, adaptarse a los posibles cambios que surgen en la investigación.

4. Métodos de investigación concluyente

4.1. Encuestas personales

La encuesta personal es el medio de obtención de información primaria más utilizado en la investigación comercial. Dada su importancia, dedicaremos el tema siguiente a analizar todos sus aspectos.

4.2. Encuestas Ómnibus

Es una técnica de recogida de información mediante entrevistas personales, por lo que tiene una gran cantidad de analogías con la entrevista personal, con la diferencia de que el cuestionario, en lugar de estar diseñado para un solo tema o producto, está compuesto de diferentes subcuestionarios relativos a otros temas o productos. Esto hace reducir el coste de la recogida de datos, ya que estará compartido por varios interesados.

Suelen realizarlos los institutos de investigación con carácter periódico, algunas veces al año, e incluso, en algunos casos, una vez cada mes.

Además, este tipo de encuesta permite ser utilizadas de forma especial para lo siguiente:

a. Para localizar subgrupos

La obtención de información de determinados subgrupos de una población a través de la encuesta personal puede ser muy costosa. Si, por ejemplo, buscamos familias que poseen una determinada marca de frigorífico, habría que descartar un elevado número de encuestados, lo que encarecería notablemente la investigación. Mediante la encuesta “ómnibus” la dificultad se reduce, ya que utiliza muestras mayores que permiten localizar un gran número de personas con las características deseadas.

b. Para acumular muestras

Al realizarse de forma periódica, permite acumular los resultados entre varias encuestas, con lo que la población encuestada es mayor, obteniéndose con ello una mayor precisión en las estimaciones correspondientes.

c. Para analizar tendencias

A través del “ómnibus” puede obtenerse la tendencia que puede tener el consumo de un determinado producto o marca a lo largo del tiempo, mediante la obtención de los datos periódicos. Por otro lado, puede conocerse la influencia de la estacionalidad en el consumo de ese producto o marca.

4.3. Experimentos

En este tipo de investigación se modifican una o más variables, por ejemplo el precio, el envase (color, tamaño, etc.), el mensaje publicitario, etc., y se observan los efectos de dichos cambios sobre otra variable, generalmente las ventas. Lo ideal es mantener constante el resto de las variables que no se han modificado, de manera que se pueda inferir que un determinado cambio en una de las variables provoca un determinado aumento en las ventas.

4.4. Observación cuantitativa

La observación cuantitativa consiste en el registro sistemático, válido y confiable de comportamiento o conducta manifiestos. Puede utilizarse como instrumento de medición en muy diversas circunstancias.

Los pasos para construir un sistema de observación son:

- Definir con precisión el universo de aspectos, eventos o conductas a observar.
- Extraer una muestra representativa de aspectos, eventos o conductas a observar.
- Establecer y definir las unidades de observación. Constituyen segmentos del contenido de las conductas o comportamientos que son caracterizados para ubicarlos en las categorías.
- Establecer y definir las categorías y subcategorías de observación. Que son los niveles o “casillas” donde serán caracterizadas las unidades de observación. Por ejemplo, en un estudio sobre el comportamiento de los consumidores, las categorías a estudiar podrían ser: “fidelidad a la marca”, “unidades de compra”, o bien, “frecuencia de compra”. Y las subcategorías consistirían en cuantificar las categorías, como por ejemplo, “alta, baja, media” o bien mediante cifras.

- **Seleccionar los observadores.** Son las personas que habrán de codificar la conducta y debe conocer las variables, categorías y subcategorías.
- **Elegir el medio de observación.** La conducta o sus manifestaciones pueden codificarse por distintos medios: observarse directamente y codificarse, o grabarse y analizarse. En algunos casos el observador se oculta y observa, en otros participa con los sujetos y codifica.
- **Elaborar las hojas de codificación.** Son las hojas donde se va a anotar las características relevantes de lo observado, para ubicarlo dentro de las categorías.
- **Proporcionar entrenamiento adecuado a los codificadores.**

La observación cuantitativa puede ser participante o no participante. En la primera, el observador interactúa con los sujetos observados, pero en la segunda no ocurre tal interacción. También puede ser indirecta cuando se recurre a los archivos que contienen los datos que nos interesa investigar.

4.5. Paneles de consumidores

El panel de consumidores es una técnica de recogida de información de carácter cuantitativo, que se realiza periódicamente a partir de una muestra permanente, representativa de la población cuya información se quiere obtener.

Los paneles básicos de consumidores existentes en el mundo pueden agruparse en dos tipos principales: panel de hogares y panel de personas que reúnen determinadas características.

4.5.1. Panel de hogares

La información recogida por el panel de consumidores depende del panel específico de que se trate. Así, por ejemplo, el panel de hogares recoge información acerca de las cantidades consumidas de los diferentes productos, las marcas y variedades, el precio unitario pagado, la forma del envase, el lugar de compra, etc.

Toda esta información es recogida en un cuestionario estandarizado que se denomina “diario de compras”, en el cual la persona objeto de estudio debe anotar diariamente los aspectos solicitados. Este diario, una vez cumplimentado, suele enviarse semanalmente por correo postal o electrónico a la

organización que controla el panel, si bien en otros casos es recogido personalmente por un empleado de esta organización.

El panel se compone de una muestra que oscila entre 3.000 y 15.000 hogares, seleccionados aleatoriamente, y que son representativos de la totalidad de los hogares existentes.

4.5.2. Panel de personas

Determinadas empresas gestionan sus propios paneles de personas, seleccionando de entre sus clientes a aquellos que cumplen determinadas características psicosociales, de acuerdo con los perfiles de sus clientes totales. Cada cierto período de tiempo solicitan de estas personas que cumplimenten una encuesta acerca de su opinión del establecimiento o la empresa, a cambio de ciertas ventajas en sus futuras compras (descuentos especiales, cheques regalo, etc.).

Dentro de los paneles de consumidores podemos encontrar las siguientes variedades:

- **El “dustbin check”:** se podría traducir como “comprobación del recipiente de los desperdicios”. Es una forma de paliar el posible olvido de anotar en el diario de compras los aspectos solicitados. Consiste en un recipiente especial en el que los componentes del panel depositan los envases y etiquetas de los productos comprados. Estos son recogidos por un empleado de la entidad que controla el panel. El inconveniente es que sólo puede realizarse para los productos que estén debidamente envasados y etiquetados.
- **El panel de audímetros:** el audímetro ha ido evolucionando a lo largo del tiempo. Actualmente, consiste en un aparato conectado a la televisión y que, por vía telefónica, envía una señal al centro controlador indicando el canal de televisión que se selecciona en cada momento. Toda esta información pasa a un ordenador central.
- **El panel de establecimientos:** está formado por una serie de establecimientos seleccionados, de los cuales se obtiene mensualmente información cuantitativa sobre diversos aspectos relacionados con la compra y venta de productos. Se efectúa mediante un inventario realizado por personal de la organización que controla el panel.

Tema 4

La encuesta y el cuestionario

I. Introducción

Los métodos de la investigación científica se desglosan de la siguiente forma:

- Métodos teóricos
- Métodos empíricos
- Métodos estadísticos matemáticos

Todos ellos están siempre relacionados, o sea, uno no puede desarrollarse sin el otro en cualquier proceso de investigación.

- **Los métodos teóricos:** Nos permiten desarrollar una teoría sobre el objeto de estudio, o sea, cómo podemos hacer una abstracción de las características y relaciones del objeto que nos expliquen los fenómenos que se investigan.
- **Los métodos empíricos:** Incluyen una serie de procedimientos prácticos sobre el objeto, que nos permiten revelar las características fundamentales y las relaciones esenciales de este, que son accesibles a la contemplación sensorial, lo cual se fundamenta en la experiencia y se expresa en un lenguaje determinado.
- **El método estadístico matemático:** Nos permite a través de tablas y cálculos matemáticos medir los resultados de los datos recopilados por medio de los instrumentales aplicados.

En cualquier investigación científica, en la etapa de recolección de datos, se usa un grupo de técnicas e instrumentos a través de los cuales podemos obtener y medir la información recopilada sobre un grupo de parámetros que queremos determinar, partiendo del diseño de la investigación, la muestra adecuada en concordancia con el problema científico a resolver y la hipótesis planteada, teniendo muy en cuenta las variables seleccionadas.

Existen varias técnicas e instrumentos para la recopilación de datos que se usan en las investigaciones científicas. En este tema nos referiremos a la técnica de la encuesta y el cuestionario como instrumento de investigación comercial.

2. La encuesta

La encuesta es una técnica de recogida de información por medio de preguntas escritas organizadas en un cuestionario impreso.

Se emplea para investigar hechos o fenómenos de forma general y no particular.

Se trata de pedir información a una muestra representativa de personas, denominados encuestados, utilizando preguntas escritas. Los cuestionarios o entrevistas recopilan datos cara a cara, teléfono, por correo o través de medios de comunicación.

Es el mejor modo de averiguar lo que el consumidor piensa. Dada la imposibilidad de tiempo y económica de entrevistar a todos los posibles miembros de la población, encuestamos solamente a una parte representativa. Nace el error de muestreo que disminuye a medida que aumenta la muestra.

En la encuesta, a diferencia de la entrevista, el encuestado lee previamente el cuestionario y lo responde por escrito, sin la intervención directa de persona alguna de las que colaboran en la investigación.

La encuesta, una vez confeccionado el cuestionario, no requiere de personal calificado a la hora de hacerla llegar al encuestado. A diferencia de la entrevista, la encuesta cuenta con una estructura lógica, rígida, que permanece inalterable a lo largo de todo el proceso investigativo. Las repuestas se recogen de modo especial y se determinan del mismo modo las posibles variantes de respuestas estándares, lo que facilita la evaluación de los resultados por métodos estadísticos.

3. Tipos de encuesta

Según su estructura, las encuestas pueden ser:

- **Estructuradas:** son aquellas en las que las preguntas y posibles respuestas están formalizadas y estandarizadas (siempre el mismo orden y de la misma forma).

- **Semi-estructuradas:** son las que presentan un guión con las principales preguntas y el orden en que deben ser formuladas, pero el orden no es estricto y el enunciado de las preguntas puede variar.
- **No estructuradas:** son aquellas que presentan un conjunto de preguntas generales sobre el tema de la investigación y tanto el orden como la forma de realizar las preguntas dependen del entrevistador y del desarrollo de la entrevista.

Según las vías de obtención de la información:

- **Directa:** Se aplica directa al sujeto.
- **Indirecta:** Se aplica por correo, teléfono, etc.

4. Fases en la elaboración de una encuesta

- **Determinación de los objetivos específicos:** en función del tipo de investigación a realizar hay que plantearse unos objetivos específicos que se pretendan conseguir con la encuesta.
- **Selección del tipo de encuesta:** determinados los objetivos, se seleccionará el tipo de encuesta a realizar, como podremos analizar en el epígrafe siguiente.
- **Diseño del cuestionario:** el diseño del cuestionario es fundamental para la recogida de datos.
- **Pilotaje del cuestionario:** es necesario probar el cuestionario previamente y marcar las instrucciones para su realización.
- **Condiciones indispensables para su realización:** en algunas ocasiones es necesario establecer las condiciones para realizar la entrevista (espacios, equipamientos, etc.).
- **Aplicación del instrumento a la muestra:** es el trabajo de campo, la realización de la entrevista.
- **Evaluación de la muestra recogida:** utilizando métodos estadísticos.

5. Formas de realizar una entrevista

5.1. Entrevista personal

Recopila información a través del contacto cara a cara con los individuos.

Ventajas

- Posible interacción para que no haya malinterpretaciones.
- Sondeo de respuestas complejas.
- Longitud de la entrevista.
- Integridad del cuestionario: reducción de las respuestas en blanco.
- Material auxiliar y ayuda visual: presentación de muestras, bocetos y otras ayudas.
- Alta participación: presencia del encuestador.

Inconvenientes

- Influencia del entrevistador: tono de voz, apariencia...
- Falta de anonimato del encuestado.
- Alto coste.

5.2. Entrevista telefónica

Recopila la información a través del contacto telefónico con los encuestados.

Ventajas

- Velocidad de recopilación de datos.
- Coste: se eliminan tiempos y gastos de desplazamiento.
- Cooperación: se coopera más en entrevistas telefónicas que por correo.

Inconvenientes

- Teléfonos móviles: aumento de hogares sin teléfono fijo y falta de ficheros con sus números de teléfono.
- Cuestionarios cortos: no sobrepasan los 5-6 minutos.
- Ausencia de ayudas visuales.

5.3. Encuesta postal

Cuestionario autoadministrado que se remite por correo al encuestado. Actualmente se sustituye esta modalidad por el correo electrónico.

Ventajas

- Flexibilidad geográfica: posibilidad de alcanzar una muestra dispersa.
- Costes: bajo coste.
- Comodidad del encuestado: según el tiempo del encuestado.
- Anonimato del encuestado: respuestas confidenciales.
- Ausencia de entrevistador: en casos de respuestas no deseadas socialmente o delicadas.
- Longitud: permite que sea un cuestionario largo.

Inconvenientes:

- Tasa de respuesta baja: en torno al 10-15%.
- Tiempo de respuesta: se alarga en el tiempo.

5.4. Encuesta por internet

Cuestionario autoadministrado emplazado en un sitio web.

Ventajas

- Velocidad y rentabilidad: gran audiencia mundial en tiempo record.
- Participación del encuestado: el usuario navega a ese sitio web para participar.
- Anonimato del encuestado: permite preguntas de tipo confidencial y delicado.

Inconvenientes

- Muestra representativa: no suele representar al total de la población especialmente si no disponen de conexión a Internet ó no están familiarizados con el ordenador y las nuevas tecnologías.

6. El cuestionario

El cuestionario es uno de los instrumentos que sirven de guía o ayuda para obtener la información deseada, sobre todo a escala masiva.

El mismo está destinado a obtener respuestas a las preguntas previamente elaboradas que son significativas para la investigación que se realiza y se aplica al universo, o a una muestra, utilizando para ello un formulario impreso, que los individuos responden por sí mismos.

Mediante el cuestionario se aspira a conocer las opiniones, las actitudes, valores y hechos respecto a un grupo de personas específico.

El cuestionario es el instrumento básico de observación en la encuesta y en la entrevista; en éste se formulan unas series de preguntas que permiten medir una o más variables, posibilitando observar los hechos a través de la valoración que hace de los mismos el encuestado o el entrevistado, limitándose la investigación a las valoraciones subjetivas de éste.

No obstante, ya que este instrumento se limita a la observación simple del entrevistador o del encuestado, puede ser masivamente aplicado a comunidades y otros grupos sociales, pudiéndose obtener información sobre una amplia gama de aspectos o problemas definidos.

7. Diseño del cuestionario

En el diseño del cuestionario es necesario iniciarse con la consigna o demanda de cooperación del entrevistado (debe realizarse de forma honesta, directa y concreta).

A la hora de elaborar un cuestionario habrá que tener en cuenta:

1º) Preguntarse lo correcto

- ¿Qué busco?
- ¿Son necesarias cada una de las preguntas?
- ¿Será suficiente una única pregunta para obtener la información?
- ¿La persona entrevistada puede realmente contestar?
- ¿La persona entrevistada aportará la información exacta?

2º) Estructurar y redactar el cuestionario

- Centrar el tema de la encuesta y orientar el cuestionario de forma precisa.
- Prever preguntas de confirmación para comprobar la fiabilidad y la coherencia de las respuestas.
- Formular preguntas que para los encuestados resulten claras, unívocas, neutras, precisas y en las que se que impliquen.

3º) Probar la eficacia del cuestionario

- Se trata de comprobar la claridad de las preguntas, su facilidad de respuesta, la duración y fluidez del cuestionario, los problemas con los que pueden tropezar los encuestadores, etc. Para ello, es útil realizar esta prueba con algunos colaboradores o conocidos.

Otras instrucciones a tener en cuenta:

- Asegure las condiciones indispensables del local.
- Preséntese oportunamente.
- Explique los propósitos del cuestionario y atienda dudas y objeciones.
- Distribuya los cuestionarios y otros materiales requeridos.
- Lea detenidamente las instrucciones y dé un ejemplo si resulta pertinente.
- Pregunte a los sujetos si han comprendido las indicaciones.
- Supervise el trabajo del grupo y auxilie a quienes lo requieran.
- Recoja los cuestionarios y dé las gracias al grupo.

Para la realización de un cuestionario eficaz y útil, siguiendo a Lilián S. Cadoche y su equipo, es necesario cumplir 17 reglas fundamentales para su elaboración:

1. Las preguntas han de ser pocas (no más de 30).
2. Las preguntas deben ser preferentemente cerradas y numéricas.
3. Redactar las preguntas con lenguaje sencillo.
4. Formular las preguntas de forma concreta y precisa.
5. Evitar utilizar palabras abstractas y ambiguas.
6. Formular las preguntas de forma neutral.
7. En las preguntas abiertas no dar ninguna opción alternativa.
8. No hacer preguntas que obliguen a esfuerzos de memoria.
9. No hacer preguntas que obliguen a consultar archivos.
10. No hacer preguntas que obliguen a cálculos numéricos complicados.
11. No hacer preguntas indiscretas.

12. Redactar las preguntas de forma personal y directa.
13. Redactar las preguntas para que se contesten de forma directa e inequívoca.
14. Que no levanten prejuicios en los encuestados.
15. Redactar las preguntas limitadas a una sola idea o referencia.
16. Evitar preguntas condicionantes que conlleven una carga emocional grande.
17. Evitar estimular una respuesta condicionada. Es el caso de preguntas que presentan varias respuestas alternativas y una de ellas va unida a un objetivo tan altruista que difícilmente puede uno negarse.

Asimismo, hay que considerar que no todas las preguntas o todas las formulaciones posibles se pueden utilizar. Consideremos algunos ejemplos de preguntas que no deben hacerse:

- **Preguntas de intelectuales:** Por ejemplo: “¿Qué aspectos particulares del actual debate positivista-interpretativo le gustaría ver reflejados en un curso de psicología del desarrollo dirigido a una audiencia de maestros?”.
- **Preguntas complejas:** Por ejemplo: “¿Cuando prepara sus clases, prefiere consultar un libro determinado incorporando la terminología que este propone o escoge varios libros de los que extrae un poco de cada uno pero que explica con sus propias palabras para hacerlos más accesibles a sus alumnos y no confundirlos?”.
- **Preguntas o instrucciones irritantes:** Por ejemplo: “¿Ha asistido alguna vez en tiempo de trabajo a un curso de cualquier clase durante su carrera entera de profesor? Si tiene más de 40 años y nunca ha asistido a un curso, ponga una marca en la casilla rotulada NUNCA y otra en la casilla rotulada VIEJO”.
- **Preguntas que emplean negaciones:** Por ejemplo: “¿Cuál es su sincera opinión sobre que ningún profesor debería dejar de realizar cursos de perfeccionamiento durante su ejercicio profesional?”.

8. Tipos de preguntas

Para la secuencia de las preguntas se debe utilizar la teoría del embudo, ir de lo más simple a lo más complejo.

En cuanto a las preguntas, habrá que observar lo siguiente:

1. Plantee preguntas que estén al nivel de conocimientos de los sujetos.
2. No utilice un lenguaje rebuscado: presente preguntas directas y sin términos de difícil comprensión.
3. Auxilie a quienes tienen dificultades para escribir sus respuestas.
4. Estimule a los sujetos que no se enfrentan con entusiasmo al cuestionario.
5. Revise los cuestionarios a la hora de recogerlos y pídale a las personas que omitieron datos que esperen durante unos minutos más.
6. No debe ser extenso en cuanto a cantidad de preguntas.

8.1. Clasificación de las preguntas

Preguntas a hacer según su función:

- **Preguntas filtro:** son aquéllas que se realizan previamente a otras para eliminar a los que no les afecte. Por ejemplo: ¿Tiene usted coche? ¿Piensa comprarse uno?
- **Preguntas trampa o de control:** son las que se utilizan para descubrir la intención con que se responde. Para ello se incluyen preguntas en diversos puntos del cuestionario que parecen independientes entre sí, pero en realidad buscan determinar la intencionalidad del encuestado al forzarlo a que las conteste coherentemente (ambas y por separado) en el caso de que sea honesto, pues de lo contrario caería en contradicciones.
- **Preguntas de introducción o rompehielos:** utilizadas para comenzar el cuestionario o para enlazar un tema con otro. Son generalmente de tipo trivial, pero sirven para comenzar la interacción entrevistador-entrevistado.
- **Preguntas muelle, colchón o amortiguadoras:** son preguntas sobre temas peligrosos o inconvenientes, formuladas suavemente.
- **Preguntas en batería:** conjunto de preguntas encadenadas unas con otras complementándose.
- **Preguntas embudo:** se empieza por cuestiones generales hasta llegar a los puntos más esenciales.

Según el grado de libertad:

- **Abierta:** En el cuestionario abierto, la persona encuestada desarrolla su respuesta, de la que el encuestador toma nota. En este caso, la encuesta de cuestionario se parece a una entrevista individual de tipo direccional. La pregunta abierta permite una respuesta libre, tanto en la forma como en la extensión.
- **Cerrada:** En el cuestionario cerrado, las preguntas marcan al encuestado una determinada forma de respuesta y una cantidad limitada de selección de respuestas. Los cuestionarios cerrados se utilizan para obtener información factual, valorar el acuerdo o el desacuerdo respecto de una propuesta, conocer la postura del encuestado respecto de una serie de juicios, etc. Pueden ser preguntas dicotómicas (sí/no), politómicas (varias opciones a elegir sólo una) o de respuesta múltiple (varias opciones y se puede elegir varias).

Según el grado de coincidencia entre objetivo y contenido:

- **Directa:** Coincide el objetivo de la pregunta con el objetivo de la investigación.
- **Indirecta:** No se corresponde el objetivo de la investigación con el contenido.

Según la correspondencia de la realidad concreta del sujeto:

- **Condicional:** Se indaga opiniones del sujeto respecto a las situaciones hipotéticas que se presentan.
- **Incondicional:** Se refiere a situaciones reales que vive el sujeto y a sus ideas, opiniones, etc.

Según el tipo de respuesta:

- **De respuesta espontánea:** son aquellas que al formularse al entrevistado no se le plantea ninguna posible alternativa, para no influirlo de ninguna forma (ejemplo: “*dígame qué marca de tabaco rubio conoce, aunque sólo sea de oídas*”).
- **De respuesta sugerida:** son las que al formularse al entrevistado se le muestra una serie de alternativas indicándole que señale aquellas que cumplen con el enunciado de la pregunta (ejemplo: “*y de las siguientes marcas de tabaco, ¿cuáles conoce?: Marlboro, Winston, Camel,...*”).

Según su contenido

- **Preguntas de identificación:** edad, sexo, profesión, nacionalidad, etc.
- **Preguntas de hecho:** referidas a acontecimientos que son concretos. Por ejemplo: ¿terminó la educación básica?
- **Preguntas de acción:** referidas a actividades de los encuestados. Por ejemplo: ¿ha realizado algún curso de capacitación?
- **Preguntas de información:** para conocer los conocimientos del encuestado. Por ejemplo: ¿sabe qué es un hipertexto?
- **Preguntas de intención:** para conocer la intención del encuestado. Por ejemplo: ¿utilizará algún programa de ordenador para controlar sus gastos?
- **Preguntas de opinión:** para conocer la opinión del encuestado. Por ejemplo: ¿qué carrera cursará después del bachillerato?

El contenido de las preguntas de una encuesta puede clasificarse de diversas maneras. La siguiente clasificación divide el campo total de los problemas en cuatro áreas amplias de contenido.

A. Datos personales:

Las encuestas a menudo incluyen preguntas relativas al sexo, la edad, la ocupación, la educación, la religión, la nacionalidad, la pertenencia a grupos y muchas otras características personales y sociales de los entrevistados. Del mismo modo, puede contener preguntas acerca del volumen de los ingresos, capitales, deudas y otras variables económicas. El propósito de estas preguntas no es tanto determinar la incidencia de estas características en la población como proporcionar las bases para el análisis de la relación del sexo, la ocupación, o los ingresos con otros datos obtenidos en encuestas.

B. Datos sobre el ambiente:

En muchas encuestas resulta importante conocer determinados hechos relacionados con las circunstancias en que viven los entrevistados. Ellos pueden abarcar datos sobre el carácter del vecindario, la adecuación de los barrios para las necesidades familiares o la proximidad de amigos o parientes.

El conocimiento de tales datos relativos al ambiente a menudo es necesario para explicar la conducta que es el principal objeto de estudio de la encuesta.

C. Datos sobre la encuesta:

Muchas preguntas de la encuesta conciernen a la conducta de los entrevistados.

Los análisis de la conducta mediante la cual obtienen información requeriría preguntas sobre lecturas de diarios, audiciones de radio y de televisión, concurrencia a cines, conversaciones y otras actividades relacionadas.

Evidentemente, el campo total de comportamiento que podría interesar al planificador de encuestas es muy amplio.

D. Nivel de información, opiniones, actitudes, motivaciones y expectativas:

Esta amplia área de datos psicológicos incluye muchas de las preguntas más interesantes que puede incluir el análisis de una encuesta. Es también el área donde es menos posible obtener datos de fuentes que no sean la encuesta.

A menudo es necesario determinar el nivel de información como un antecedente para el estudio de actitudes u opiniones. Es peligroso suponer que todos comprenden en igual medida los temas y los acontecimientos y es difícil estimar la posición de las personas sin conocer su comprensión del asunto en cuestión. El nivel de información de un entrevistado puede medirse simplemente en términos de si conoce o no un problema o hecho.

Las actitudes son puntos de vista generalizados de aprobación o desaprobación. Determinar la presencia o ausencia de actitudes y las razones que las justifican —*¿qué temas públicos son aprobados o rechazados, por qué clase de personas y por qué motivos?*— es con frecuencia un importante objetivo de la encuesta.

Habitualmente, los analistas de encuestas no se satisfacen con obtener información sobre actitudes respecto de temas públicos específicos y no relacionados entre sí; a menudo consideran más interesante estudiar pautas de actitudes e interrelaciones entre diferentes actitudes. De esta manera, pueden discernirse actitudes generales y es posible determinar qué actitudes específicas son influidas y cuáles no por estas actitudes generales.

El estudio de las motivaciones y expectativas representa una de las áreas más interesantes de la investigación por encuestas. El concepto de “motivación” no sólo se refiere a las razones declaradas de la conducta —es decir las respuestas a las preguntas “por qué”—, sino, en un sentido más general, a las fuerzas que impulsan la acción. Las expectativas representan la perspectiva temporal de una persona como proyección hacia el futuro; es decir, sus opi-

niones y actitudes acerca de lo que ocurrirá, como también sus intenciones y planes.

9. El control de los evaluadores y los gestores

El resultado de la encuesta será más o menos válido en función del rigor que haya seguido su elaboración, la recogida de datos y su evaluación posterior. Por ello, para determinar estos aspectos, podemos seguir listas de control para cerciorarnos del cumplimiento o no de los requisitos fundamentales de obtención de datos.

Estas listas de control pueden establecerse tanto para los evaluadores como para los gestores del estudio realizado.

Listas de control para evaluadores:

- ¿Se ha justificado la realización de un cuestionario cerrado con una muestra representativa por la necesidad de obtener indicadores estadísticos?
- ¿Se ha realizado la encuesta entre una muestra representativa?
- ¿Son claras y sencillas las preguntas realizadas? ¿Y las respuestas?
- ¿La extensión del cuestionario ha sido adaptada?
- ¿Incluye el cuestionario preguntas cruzadas o de control?
- ¿Se ha probado el cuestionario?
- ¿Se ajustan las técnicas de realización del cuestionario al tipo de público encuestado (cara a cara, teléfono, etc.)?
- ¿Se ha puesto en marcha un mecanismo de seguimiento y de control de los encuestadores?
- ¿Se han organizado reuniones de formación/contextualización?
- ¿Son los encuestadores independientes respecto de la política o el programa evaluados?
- ¿Es el número de encuestados suficiente como para ser representativo?
- ¿Son los resultados cuantitativos lo bastante precisos como para ser utilizados en la evaluación?

- ¿Se presentan y especifican los resultados según las diversas categorías de actores y beneficiarios?
- ¿Se han articulado los resultados con las demás herramientas de información y de análisis empleadas por los evaluadores?

Lista de control para gestores:

- ¿Se ha justificado de forma clara el empleo de la herramienta?
- ¿Se ha realizado la encuesta entre una muestra representativa?
- ¿Se ha probado el cuestionario?
- ¿Ha permitido el tratamiento del cuestionario obtener los indicadores deseados y la suficiente precisión?
- ¿Se han articulado los resultados con las demás herramientas de información y de análisis empleadas por los evaluadores?

En la siguiente página se presenta un modelo de cuestionario.

Cuestionario nº: _____

Buenos días/tardes, estamos realizando un estudio sobre móviles, los hábitos y los gustos de los consumidores. Nos gustaría contar con su opinión para desarrollar el estudio y le garantizamos el total anonimato de sus respuestas que serán tratadas de forma anónima con las del resto de los participantes. Solo le llevará unos minutos. Gracias

1. ¿Tiene Vd. teléfono móvil?

1. ☐ Sí

2. ☐ No (GRACIAS, NO VALE)

2. ¿Cuántos años hace que tiene Vd. móvil?

3. ¿Cuánto paga usted aprox. cada mes por telefonía móvil?

4. ¿Cuántas llamadas suele Vd. realizar al día?

5. ¿Cuál suele ser el horario de sus llamadas?

☐ Noche

☐ Mañana

☐ Tarde

6. ¿Cuántas llamadas realiza cada día de las que le indico a continuación y en qué horario?

	Número de llamadas	Horario (mañana, tarde, noche)
Personales		
Profesionales		
Mixtas		
A números 900		

7. ¿Podría indicarnos cuál es su actual compañía de telefonía móvil?

(Marcar con una x solo una respuesta en P7)

8. ¿Qué compañías de telefonía móvil conoce, además de la suya?

(Marcar con una x las respuestas en P8)

9. ¿Podría clasificar de 1 a 5 las siguientes compañías de telefonía móvil según sus preferencias, siendo 5 la mejor y 1 la peor?

(Marcar de 1 a 5 las respuestas en P9)

	P7	P8	P9
Vodafone			
Orange			
Movistar			
Yoigo			
Simyo			

10. ¿En que medida influyen las siguientes aspectos en su elección de la compañía de telefonía móvil? (Escriba el número, siendo 1 poco importante y 5 muy importante)

Atención al cliente por vía telefónica	
Atención al cliente en establecimientos	
Solución de problemas	
Simpatía de los operadores	
Ofertas especiales	
Cobertura	
Claridad de la señal	
Servicios adicionales	
Programa de puntos	
Tarifas	
Cuota de mercado	
Ser la compañía de mis familiares o amigos	
Aviso de novedades	
Página web	
Promociones	

11. ¿Tiene usted algún otro motivo que considere importante para elegir su compañía?

12. Valore la importancia de los siguientes tipos de servicios de telefonía móvil (Escriba el número, siendo 1 poco importante y 5 muy importante)

Llamadas de voz	
Buzón de voz	
SMS	
GPS	
Internet	
Música	

13. ¿Utiliza Vd. el servicio de buzón de voz?

☐ 1 Sí

☐ 2 No (pasar a pregunta 17)

14. ¿Cuántas veces utiliza Vd. cada día el servicio de buzón de voz?

15. Valore las siguientes características del servicio de buzón de voz (Siendo 1 poco importante y 5 muy importante)

Claridad de sonido	1	2	3	4	5
Facilidad de acceso	1	2	3	4	5
Coste	1	2	3	4	5
Seguridad	1	2	3	4	5
Aviso en el móvil	1	2	3	4	5

16. ¿Tiene usted alguna otra característica del servicio de buzón de voz que considere importante?
- _____

PREGUNTAS DE CLASIFICACIÓN

17. Edad: _____ años

18. Estado civil:

Solter@	☐ 1	Separad@	☐ 4
Casad@	☐ 2	Divorciad@	☐ 5
Con pareja	☐ 3	Viud@	☐ 6

19. Lugar de residencia:

Valencia Ciudad ☐ 1 Otros ☐ 2

20. Nivel de estudios del cabeza de familia (*principal perceptor de ingresos*):

Menos primaria no lee	☐ 1	Bachill., COU, FP2	☐ 5
Primarios incompletos	☐ 2	Diplomado	☐ 6
Primarios	☐ 3	Licenciado	☐ 7
ESO, BUP, FPI	☐ 4		

21. Nivel de ingresos medios mensuales de la unidad familiar:

Hasta 600 €	☐ 1	De 2.101 a 2.500 €	☐ 5
De 601 a 900 €	☐ 2	De 2.501 a 3.000 €	☐ 6
De 901 a 1.500 €	☐ 3	De 3.001 a 4.000 €	☐ 7
De 1.501 a 2.100 €	☐ 4	Más de 4.000 € al mes	☐ 8

22. Sexo

☐ 1 hombre ☐ 2 mujer

MUCHAS GRACIAS POR SU COLABORACIÓN

Entrevistador: _____

Zona de entrevista: _____

Fecha entrevista: _____

Tema 5

La tabulación de encuestas

1. Las variables estadísticas

En los estudios estadísticos se busca investigar acerca de una o varias características de la población observada. Para un correcto manejo de la información, estas características deben ser tomadas en cuenta de acuerdo a su tipo para poder hablar de la aplicación de algunas de las operaciones que más adelante se llevarán a cabo.

Una variable es una función que asocia a cada elemento de la población la medición de una característica, particularmente de la que se desea observar.

De acuerdo a la característica que se desea estudiar, a los valores que toma la variable, se tiene la siguiente clasificación:

- **Variables categóricas:** son aquellas cuyos valores son del tipo categórico, es decir, que indican categorías o son etiquetas alfanuméricas o “nombres”. A su vez se clasifican en:
 - *Variables categóricas nominales:* son las variables categóricas que, además de que sus posibles valores son mutuamente excluyentes entre sí, no tienen alguna forma “natural” de ordenación. Por ejemplo, cuando sus posibles valores son: “sí” y “no”. A este tipo de variable le corresponde las escalas de medición nominal.
 - *Variables categóricas ordinales:* son las variables categóricas que tienen algún orden. Por ejemplo, cuando sus posibles valores son: “nunca sucede”, “la mitad de las veces” y “siempre sucede”. A este tipo de variable le corresponde las escalas de medición ordinal.
- **Variables numéricas:** toman valores numéricos. A estas variables le corresponde las escalas de medición de intervalo, y a su vez se clasifican en:

- *Variables numéricas discretas*: son las variables que únicamente toman valores enteros o numéricamente fijos. Por ejemplo: las ocasiones en que ocurre un suceso, la cantidad de euros que se gastan en una semana, los barriles de petróleo producidos por un determinado país, los puntos con que cierra diariamente una bolsa de valores, etc.
- *Variables numéricas continuas*: llamadas también variables de medición, son aquellas que toman cualquier valor numérico, ya sea entero, fraccionario o, incluso, irracional. Este tipo de variable se obtiene principalmente, como dice su nombre alterno, a través de mediciones y está sujeto a la precisión de los instrumentos de medición. Por ejemplo: el tiempo en que un corredor tarda en recorrer una cierta distancia (depende de la precisión del cronómetro usado), la estatura de los alumnos de una clase (depende de la precisión del instrumento para medir), la cantidad exacta que despacha un surtidor de combustible en una gasolinera (para efectos de regulación y fiscalización, y depende de la precisión del instrumento para medir volúmenes), etc.

2. Tabulación de encuestas

La tabulación de encuestas son las distintas formas de medir o cuantificar las respuestas a determinadas preguntas, principalmente aquellas relacionadas con sentimientos, actitudes, opiniones y creencias. Nos centraremos fundamentalmente en estas, puesto que las variables numéricas ya están cuantificadas por definición.

Cuantificar y medir las respuestas nos permite:

- Sintetizar la información.
- Aplicar técnicas que permiten enriquecer esa información.

De entre las múltiples técnicas de tabulación de datos ofrecemos a continuación las más utilizadas, agrupadas en:

- Escalas básicas.
- Escalas comparativas.
- Escalas no comparativas.

2.1. Escalas básicas

Son el punto de partida de creación del resto de escalas. Desde la nominal a la de ratios, cada una de ellas ofrece mayor precisión en la medición, desde

el punto de vista estadístico, y en el uso posterior de la información.

Dentro de este grupo encontramos las siguientes escalas:

Escala nominal: se utiliza únicamente para identificar diferentes categorías o alternativas de respuesta. La asignación de valores a las distintas respuestas es arbitraria: no encierran ningún significado ni indican orden alguno.

Ejemplo: ¿Cuál de estas marcas de automóviles conoce?

- 1 Marca A
- 2 Marca B
- 3 Marca C
- 4 Ninguna de ellas

Escala ordinal: asigna diferentes valores a distintas respuestas con la intención de asignar un rango u orden. La diferencia entre los intervalos no tiene ningún significado.

Ejemplo: Ordene de mayor a menor las siguientes marcas de automóviles según su preferencia.

- 1 Marca A
- 2 Marca B
- 3 Marca C

Escala de intervalo: presenta distintas alternativas de respuesta con números asociados. Estos números muestran un orden y además la diferencia entre los valores de la escala es constante y posee significado (el valor que separa una respuesta concreta con la contigua es el mismo que separa esta última con la siguiente).

Ejemplo: ¿Cambiará Vd. de automóvil en los próximos 12 meses?

- 1 Seguro que no
- 2 No es muy probable
- 3 Es probable
- 4 Seguro que sí

Escala de ratio: tienen las características de las anteriores y además permiten la obtención de ratios coherentes con sus valores. Se conoce perfectamente el punto de origen (cero euros) pudiendo realizarse comparaciones

con las distintas respuestas. Si un individuo gana 2.000,00 € al mes gana el doble que otro de 1.000,00 € al mes.

Ejemplo: ¿Cuánto dinero gasta mensualmente en sus compras en supermercado?

Euros

2.2. Escalas comparativas

Son un conjunto de escalas en las que las valoraciones se llevan a cabo de forma relativa, atendiendo a un elemento de referencia (conjunto a comparar), es decir, aportan al individuo un punto de referencia a la hora de elaborar su juicio.

Resultan convenientes cuando el individuo no tiene conocimiento o experiencia de la cuestión planteada por el investigador.

Dentro de las escalas comparativas podemos mencionar las siguientes:

Escala de comparaciones pareadas: se basan en la presentación de los estímulos o elementos a comparar por pares, simplificándose al máximo cada una de las elecciones. Es muy utilizada para evaluar productos ya existentes en el mercado, respecto a los demás.

Ejemplo: De las siguientes parejas de marcas de automóviles señale la que prefiere en cada caso

SEAT – OPEL

SEAT – RENAULT

OPEL – RENAULT

Una vez tengamos las respuestas habrá que transformarlas en escala ordinal determinando el orden de preferencia del individuo. Así, la marca más deseada será aquella que ha elegido más veces, la segunda será la siguiente, etc.

Escala de clasificación: se basa en pedir al entrevistado que ordene o clasifique un conjunto de estímulos en función de un atributo.

Ejemplo: Clasifique según su opinión las siguientes marcas de automóviles según el grado de seguridad que ofrecen, siendo 1 la que más seguridad tiene y 4 la que menos.

Marca A ____

Marca B ____

Marca C ____

Marca D ____

Escala de suma constante: se utiliza para medir la importancia relativa que el entrevistado asigna a los estímulos o variables, ya que se le pide que reparta una cantidad de puntos fija (generalmente 100) entre los mismos. El gran inconveniente es que hay que realizar cálculos para contestar.

Ejemplo: Reparta 100 puntos entre las características siguientes de forma que refleje cuál es la importancia que tiene para Vd. cada una de ellas a la hora de optar por la compra de un automóvil.

Precio _____

Motor _____

Potencia _____

Consumo _____

Diseño _____

Escala de Guttman: se trata de un tipo de escala que ordena todas las respuestas en base a una sola característica o atributo, presentándose los estímulos de sencillos a más complejos. Puede sustituir a un conjunto de preguntas dicotómicas, en las que una respuesta afirmativa a una de las mismas implica una respuesta afirmativa a las anteriores.

Ejemplo: Señale los estudios que ha alcanzado:

- 1 Ninguno
- 2 Ninguno, pero sabe leer
- 3 Estudios primarios
- 4 Estudios secundarios
- 5 Bachillerato
- 6 Ciclo formativo superior
- 7 Diplomatura o ingeniería técnica
- 8 Universitarios superiores
- 9 Doctorado

Escala de clases o similitudes: son usadas para clasificar a un número elevado de estímulos en un número de subconjuntos o grupos reducidos, atendiendo a la similitud de los mismos. Suele usarse como paso previo a una clasificación ordinal.

Ejemplo: De los siguientes automóviles, clasifique cada uno en alguno de los siguientes grupos:

Marca A	Marca B	Marca C	Marca D	Marca E

No tiene por qué coincidir el número de elementos a clasificar con el número de grupos de clasificación.

Escala de protocolos verbales: es un tipo de escala en la que se pide la opinión al entrevistado frente al estímulo planteado, mostrándose las posibles respuestas en forma de enunciados verbales.

Ejemplo: ¿Qué le parece la marca de detergente A en comparación con el que usa habitualmente?

- ☐ Es muchísimo mejor
- ☐ Es mejor
- ☐ Es más o menos igual
- ☐ Es peor
- ☐ Es muchísimo peor

Utilizar o no un número par de alternativas permitirá o no una respuesta neutral o intermedia.

2.3. Escalas no comparativas

En este tipo de escalas, las contestaciones a los distintos enunciados no se fundamentan en la comparación entre estímulos o variables.

Se suele utilizar para medir valoraciones personales. Dentro de ellas podemos analizar las siguientes:

Escalas de clasificación continua: están diseñadas para medir la opinión de los entrevistados, presentando un elevado número de alternativas de respuesta a través de una línea continua. También puede usarse para clasificaciones numéricas.

Ejemplo: Indique por favor, marcando sobre esta línea, cuál es su opinión respecto a la amabilidad en el trato recibido por el vendedor en esta tienda.

Nada amables _____ amables
0 10

Presentan el gran inconveniente de su dificultad de codificación y medición de las respuestas.

Escalas de Likert: se emplea habitualmente para medir actitudes. Consiste en crear un conjunto de enunciados para que el entrevistado muestre su nivel de acuerdo o desacuerdo. Una vez asignados los valores a las distintas declaraciones habrá que sumar las puntuaciones que se han dado al total de todas las declaraciones.

Ejemplo: Por favor, indique su grado de satisfacción con el servicio recibido en esta tienda.

	Muy insatisfecho	Insatisfecho	Mediocre	Satisfecho	Muy satisfecho
Trato recibido	-2	-1	0	1	2
Soluciones aportadas	-2	-1	0	1	2
Ofertas recibidas	-2	-1	0	1	2
Confiabilidad	-2	-1	0	1	2

En este ejemplo, cada declaración variará entre -8 y +8 (ya que hay 4 declaraciones). Dicha puntuación permitirá valorar la actitud general del individuo hacia el servicio recibido.

Escalas de diferencial semántico: evalúan un estímulo en función de diversos atributos, adjetivos o sentencias bipolares, separados por 7 categorías de respuesta. Se analizan tanto las puntuaciones totales como los perfiles obtenidos. Puede usarse para analizar y comparar diversos estímulos de forma simultánea.

Ejemplo: ¿Cuál es su opinión sobre los *banner* como publicidad en televisión?

	1	2	3	4	5	6	7	
De ningún interés								De gran interés
Nada creíbles								Muy creíbles
No impresionan								Impresionan mucho
Nada atractivos								Muy atractivos
Nada informativos								Muy informativos
Nada claros								Muy claros
No llaman la atención								Llaman la atención
No gustan								Gustan mucho
Nada convincentes								Muy convincentes

De esta forma, estamos analizando atributos como: interés, credibilidad, atractivo, etc.

Tema 6

El muestreo

1. Teoría del muestreo

Una vez realizado el cuestionario, el siguiente paso es determinar el universo sobre el que ha de investigarse. Consiste en acotar la población a la que se dirigirá el estudio dependiendo del enfoque de la investigación comercial.

Cuando se trabaja con universos muy numerosos, resulta imposible entrevistar a todos. Para resolver el problema se emplea la teoría del muestreo. Esta teoría nos permite conocer aspectos del universo a través de una pequeña muestra del mismo. La estadística responde a esta suposición con la llamada ley de los grandes números. Según Bernouilli, cualquiera que sea el grupo de objetos, extraído de otro grupo más importante, tenderá a presentar las mismas características que el grupo mayor.

Recibe el nombre de muestra una parte del grupo de elementos que se examinan y población es la totalidad de los sucesos de un fenómeno dado o del grupo de elementos a estudiar; estos pueden ser cualquier cosa que se pueda medir, contar o jerarquizar.

Hay que distinguir ahora entre población finita, que es aquella que tiene un número limitado de elementos o sucesos, y la población infinita, que cuenta con un número infinitamente grande de elementos.

2. Tipos de muestreo

En primer lugar, cabría distinguir entre muestreo probabilístico y no probabilístico.

El muestreo no probabilístico es aquel en el que el analista elige, según su propio criterio, los individuos de la muestra. La selección puede realizarse por cuotas representativas de la población (porcentajes determinados para cada característica psicográfica, por ejemplo). Otra forma de muestreo no

probabilístico sería el denominado “bola de nieve”, donde se solicita a los propios seleccionados que identifiquen nuevos individuos para añadir a la muestra.

El muestreo probabilístico pretende fijar, mediante probabilidades, la muestra a elegir. Entre los diversos tipos de muestreo probabilístico podemos analizar los siguientes:

2.1. Muestreo aleatorio simple

Entre la gran variedad de métodos que existen para elegir una muestra, quizás el más importante sea el muestreo aleatorio simple. En términos generales, es un procedimiento de elegir una muestra de la población de forma que tenga la misma probabilidad de ser seleccionada que las demás.

Se trata de sortear entre todos los componentes del universo, aquellas personas que van a formar parte de la muestra calculada. Así, se obtiene la mayor representatividad posible. Pero en la práctica este método presenta dificultades:

- Imposibilidad de poder relacionar a todo el universo en una lista.
- Coste y confección de esas listas.
- Inconvenientes que presenta la obligada entrevista a las personas seleccionadas.
- Una forma de simplificar el método sería la utilización de rutas aleatorias (se sortean zonas geográficas).

2.2. Muestreo sistemático

Según este procedimiento, se obtiene una muestra tomando cada k-ésima unidad de la población tras numerar los elementos de está o de ordenarlos de alguna manera. La letra k representa un número entero, que es aproximadamente la razón de muestreo entre el tamaño de la población y el tamaño de la muestra. Así, si la población consiste en 1.000 elementos de muestreo y se desea una muestra de 50 elementos, entonces

$$K = 1000/50 = 20$$

La muestra se obtiene tomando una unidad cada veinte de la población. También se puede emplear fácilmente cuando se dispone ya de una lista de los elementos de la población.

El procedimiento del muestreo aleatorio sistemático selecciona una muestra más representativa que la del muestreo aleatorio simple si los elementos cercanos de la población se parecen más entre sí, y es menos representativa en situaciones en las que hay periodicidad oculta en la población.

2.3. Muestreo por estratos

Para el muestreo restringido se entiende que el muestreo aleatorio por estratos es más efectivo que el método simple, ya que este procedimiento exige tener un conocimiento previo de la población. El procedimiento de estratificación contempla el dividir la población en grupos llamados estratos; dentro de cada uno de ellos están los elementos situados de manera más homogénea en base a las características que se van a estudiar. El razonamiento es que mediante este ordenamiento, la variabilidad es menor que la de la población total y por tanto se necesita un tamaño de muestra más pequeño.

Por ejemplo, cada estrato podría ser un conjunto de nivel cultural, raza y edad. El inconveniente es que exige un conocimiento elevado de la población a estudiar.

2.4. Muestreo por conglomerados

Este método consiste en seleccionar primero al azar grupos, llamados conglomerados, de elementos individuales de la población, y en tomar luego todos los elementos o una submuestra dentro de cada conglomerado para constituir así la muestra global. Sería preferible que cada conglomerado fuese una miniatura de toda la población y así se obtendría una muestra satisfactoria. Pero en la práctica, este requisito es muy difícil de llevar a cabo.

Los conglomerados suelen ser llamados también unidades de muestreo primarios. Si todos los elementos seleccionados de estos se incluyen en la muestra, el procedimiento se denomina muestreo de una etapa pero si se saca una submuestra aleatoria de elementos de cada conglomerado seleccionado se llama muestreo de dos etapas y si intervienen más de dos etapas en la obtención de la muestra global, se dice que es un muestreo de etapas múltiples.

El objetivo del muestreo por conglomerados es el estudio de las características de los elementos individuales, si bien se eligen inicialmente las unidades de muestreo primarias. La ventaja principal de este muestreo es la gran

reducción de costes con un grado dado de fiabilidad; además es utilizado a menudo en el control de calidad estadístico y la desventaja está en su relativa ausencia de fiabilidad en un tamaño dado de la muestra.

Ejemplo: para realizar una encuesta a los propietarios de restaurantes en una ciudad se sortean de forma aleatoria las calles de esa ciudad y se cuenta el número de restaurantes de cada una, hasta llegar al tamaño de la muestra que se desea (1ª calle, 3 restaurantes; 2ª calle, 5 restaurantes; etc. hasta llegar al número de restaurantes que se quiere entrevistar).

3. Tamaño de la muestra y error de muestreo

En primer lugar, cabe determinar el tamaño de la población, distinguiéndose entre poblaciones finitas e infinitas. En general, se acepta que:

Población finita ≤ 100.000 elementos

Población infinita > 100.000 elementos

Para determinar el tamaño de la muestra necesitamos definir una serie de parámetros:

N = tamaño de la población

n = tamaño de la muestra

K = error de muestreo

$p(1-p)$ = dispersión

S = desviación típica

El valor de “ p ” se puede obtener de los resultados de un pre-test, pero la práctica habitual es suponer la dispersión máxima, o sea, con $p = 0,5$. Esto no implica un incremento relevante en el tamaño de la muestra y es en realidad la opción menos arriesgada.

La desviación típica (S) se calcula para poblaciones finitas con la siguiente fórmula:

$$S = \sqrt{\frac{(N-n) p(1-p)}{(N-1) n}}$$

Y para poblaciones infinitas, con la fórmula siguiente:

$$S = \sqrt{\frac{p(1-p)}{n}}$$

Se llama intervalo de confianza a un par de números entre los cuales se estima que estará cierto valor desconocido con una determinada probabilidad de acierto. Formalmente, estos números determinan un intervalo, que se calcula a partir de datos de una muestra, y el valor desconocido es un parámetro poblacional. La probabilidad de éxito en la estimación se denomina nivel de confianza.

El nivel de confianza y la amplitud del intervalo varían conjuntamente, de forma que un intervalo más amplio tendrá más posibilidades de acierto (mayor nivel de confianza), mientras que para un intervalo más pequeño, que ofrece una estimación más precisa, aumentan sus posibilidades de error.

El error muestral (K) depende del grado de confianza admitido. A mayor nivel de confianza la estimación es menos precisa y por tanto habrá mayor error muestral. De esta forma, tenemos lo siguiente:

Nivel de confianza del 68%: $K = 1S$

Nivel de confianza del 95%: $K = 2S$

Nivel de confianza del 99%: $K = 3S$

Generalmente, se suele trabajar con un nivel de confianza del 95% y por lo tanto, con un error muestral de $2S$.

El tamaño de la muestra (n), con un nivel de confianza del 95% dependerá de si las poblaciones son finitas o infinitas. Así, tenemos que:

a) Para poblaciones finitas:

$$n = \frac{4Np(1-p)}{k^2(N-1) + 4p(1-p)}$$

b) Para poblaciones infinitas:

$$n = \frac{4p(1-p)}{k^2}$$

Ejemplo 1: Determinar el tamaño óptimo de la muestra para una encuesta a jóvenes de entre 15 y 18 años en España con un nivel de confianza del 95%, asumiendo un error del 5%, sabiendo que, según datos del INE (Instituto Nacional de Estadística) el total de jóvenes de esas edades es de 2 millones.

Solución 1: Estamos ante una población infinita al ser mayor de 100.000 elementos. El valor de K es 0,05 y p lo consideraremos siempre de 0,5.

$$n = \frac{4p(1-p)}{k^2}$$

$$n = \frac{4 \cdot 0,5 (1-0,5)}{0,05^2} = 400 \text{ unidades muestrales}$$

Como el error muestral es del 5%, podemos decir que con un grado de confianza del 95% cualquier dato obtenido de la muestra se ajustará a la realidad de la población en $\pm 5\%$. Así, por ejemplo, si obtenemos como media de gastos mensuales 100,00 €, podemos decir que con un 95% de confianza, la media de gasto de la población de jóvenes entre 15 y 18 años en España es de 95,00 a 105,00 €.

Ejemplo 2: Estimar el error de muestreo cometido al analizar una muestra de 1.000 individuos dentro de una población de 10.000, con un nivel de confianza del 95%.

Solución 2: Sabemos que p es 0,5 y que la población es finita (inferior a 100.000 elementos). Con un nivel de confianza del 95%, sabemos también que $K = 2S$. Por tanto, calcularemos S con la fórmula correspondiente:

$$S = \sqrt{\frac{(N-n) p(1-p)}{(N-1) n}}$$

$$S = \sqrt{\frac{(10.000 - 1.000) 0,5(1-0,5)}{(10.000 - 1) 1.000}} = 0,015$$

Por tanto, el error de muestreo es:

$$K = 2S$$

$$K = 2 \cdot 0,015 = 0,03$$

El error cometido es del 3%. Por tanto podemos decir que con un grado de confianza del 95% cualquier dato obtenido de la muestra se ajustará a la realidad de la población en $\pm 3\%$. Así, por ejemplo, si obtenemos como media de sueldo mensual 1.000,00 €, podemos decir que con un 95% de confianza, la media de gasto de la población está entre 970,00 y 1.030,00 €.

Ejercicios tema 6

1. Vamos a realizar una encuesta en una zona de 25.000 habitantes y queremos elegir una muestra con un grado de confianza del 95%, aceptando un error muestral del 2%.
 - a) ¿Cuál debe ser el tamaño de la muestra?
 - b) Si, debido al presupuesto disponible, decidimos encuestar a 3.000 personas ¿cuál será el error muestral?
 - c) Encuestando a 3.000 personas hemos obtenido que el 12% son menores de 15 años ¿cómo podemos extrapolar esta información a la población? ¿cuántos menores de 15 años podemos estimar que hay en la población?
 - d) Disponemos del censo de la población y elegiremos la muestra de 3.000 personas según el método sistemático ¿cómo realizaremos la selección?
2. Vamos a realizar una encuesta en una ciudad de 250.000 habitantes y queremos elegir una muestra con un grado de confianza del 95%, aceptando un error muestral del 3%.
 - a) ¿Cuál debe ser el tamaño de la muestra?
 - b) Si, debido al presupuesto disponible, decidimos encuestar a 5.000 personas ¿cuál será el error muestral?
 - c) Encuestando a 5.000 personas hemos obtenido que el 25% son fumadores ¿cómo podemos extrapolar esta información a la población? ¿cuántos fumadores podemos estimar que hay en la población?
 - d) Elegiremos la muestra de 5.000 personas según el método por conglomerados ¿cómo realizaremos la selección?

Tema 7

Estadísticas descriptivas simples

1. Introducción a la Estadística

La Estadística estudia los fenómenos de masa para hallar en ellos las regularidades del comportamiento colectivo, regularidades que sirven para describir el fenómeno y para efectuar predicciones.

Es obvio, pues, que toda investigación estadística ha de estar necesariamente referida a un conjunto o colección de personas o cosas. A este conjunto se le denomina técnicamente “población”. Los elementos de los que se compone dicha población poseen ciertas propiedades, rasgos, o cualidades que se denominan “caracteres”.

Estos caracteres pueden ser:

- **Cuantitativos:** también llamados variables y son los que se describen mediante números (edad, altura, ingresos, etc.).
- **Cualitativos:** también llamados atributos y son los descritos por palabras (profesión, nacionalidad, estado civil, etc.).

Llamaremos tamaño de una población al número de elementos que la componen. Normalmente, en la ciencia estadística se recurre al estudio de grupos de elementos de tamaño más o menos reducido, que se conoce con el nombre de muestra. Una muestra no es más que una parte de la población que sirve para representarla.

2. Clasificación de la Estadística

Según los objetivos que se pretendan, la Estadística puede dividirse en:

- **Estadística descriptiva:** es la ciencia dedicada a descubrir las regularidades o características existentes en un conjunto de datos.

- **Estadística inductiva o inferencia estadística:** tiene como función generalizar los resultados de una muestra para estimar las características de la población.

Según el número de caracteres investigados, los estudios estadísticos pueden clasificarse en:

- **Estadísticas simples:** estudian el comportamiento de un único carácter.
- **Estadísticas múltiples:** estudian el comportamiento de varios caracteres simultáneamente.

3. Tipos de presentación de datos en Estadística de una variable

En la práctica, la manera de obtener una Estadística de una sola variable depende de los siguientes factores:

- Del número de observaciones efectuadas.
- Del número de valores distintos que tome la variable.

De la conjunción de los dos factores mencionados resultarán tres tipos de presentación de datos:

TIPO I: son estadísticas que constan de pocas observaciones. La forma de presentarlas es simplemente anotando las observaciones en fila o en columna de menor a mayor. Ejemplo: análisis de alturas en 6 jóvenes:

ALTURAS	
1.58 cm	
1.60 cm	
1.61 cm	
1.65 cm	
1.66 cm	
1.72 cm	

TIPO II: son estadísticas que constan de muchas observaciones, pero la variable toma pocos valores distintos. Se presentan los datos anotando en una primera columna los valores distintos de la variable, ordenados de menor a mayor, y en una segunda columna, correspondiéndose con la pri-

mera, la frecuencia o número de veces que cada valor aparece repetido. Esta frecuencia se denomina absoluta. La variable se denota por X y cada uno de sus valores serán $x_1, x_2, \dots, x_k$. Las frecuencias absolutas se denotan por $n_1, n_2, \dots, n_k$. La suma total de las frecuencias absolutas será el número total de observaciones y se denota por N . Ejemplo: análisis de número de hijos en una muestra de 100 parejas (o sea, $N=100$):

HIJOS = X_i	FRECUENCIA = n_i
0	10
1	40
2	30
3	15
4	5

TIPO III: son estadísticas que constan de muchas observaciones y la variable toma muchos valores distintos. Se representa agrupando en una cuantas clases todos los valores distintos de la variable. Cada clase será, por tanto, un intervalo cuyos límites denotaremos por $[L_{i-1} , L_i)$. Una vez definidas las clases se procede a obtener la frecuencia de cada clase que estará constituida por el número de valores que se encuentra en cada intervalo. La amplitud de cada intervalo viene definida por la diferencia $L_i - L_{i-1}$. Esta amplitud puede permanecer constante en todos los intervalos o ser variable. Para realizar los cálculos estadísticos con los valores de la variable se utilizarán los puntos medios de cada clase, llamados marcas de clase, que se denotan por $x_1, x_2, \dots, x_k$ y se calculan $(L_i + L_{i-1})/2$. Ejemplo: análisis de ingresos anuales en una muestra de 150 personas (o sea, $N=150$):

CLASES= $[L_{i-1} , L_i)$	MARCA DE CLASE= X_i	FRECUENCIAS= n_i
0 – 6.000	3.000	4
6.000 – 12.000	9.000	45
12.000 – 18.000	15.000	66
18.000 – 24.000	21.000	25
24.000 – 35.000	29.500	9
35.000 – 60.000	47.500	1

4. Distribución de frecuencias

La presentación de los datos para su estudio estadístico se realiza con las tablas antes referidas y que reciben el nombre de distribuciones de frecuencias. En ellas pueden aparecer además:

- **Frecuencias absolutas acumuladas:** se denotan por $N_1, N_2, \dots, N_k$ y son la suma de todas las frecuencias absolutas desde el primer valor hasta el correspondiente.
- **Frecuencias relativas:** se simboliza por $f_1, f_2, \dots, f_k$ y es el cociente entre la frecuencia absoluta y el número de elementos que se observan (N). Se calcula $f_i = n_i / N$. Indican la proporción que, dentro del total, supone un valor concreto de la variable. La suma de todas las frecuencias relativas es, evidentemente, 1.
- **Frecuencias relativas acumuladas:** se simbolizan por $F_1, F_2, \dots, F_k$ y se obtienen acumulando las frecuencias relativas.

5. Representaciones gráficas

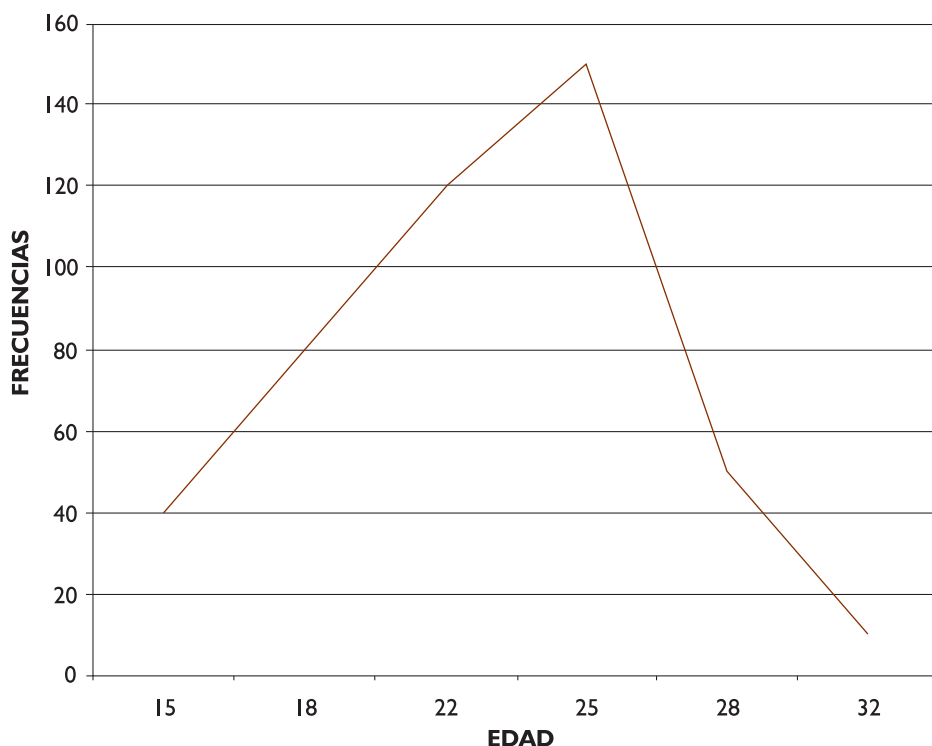
La finalidad de representar gráficamente los datos contenidos en una tabla estadística es ofrecer una visión de conjunto del fenómeno sometido a investigación, más rápidamente perceptible que la observación directa de los datos numéricos.

Las representaciones gráficas más usuales para estadísticas de una variable son:

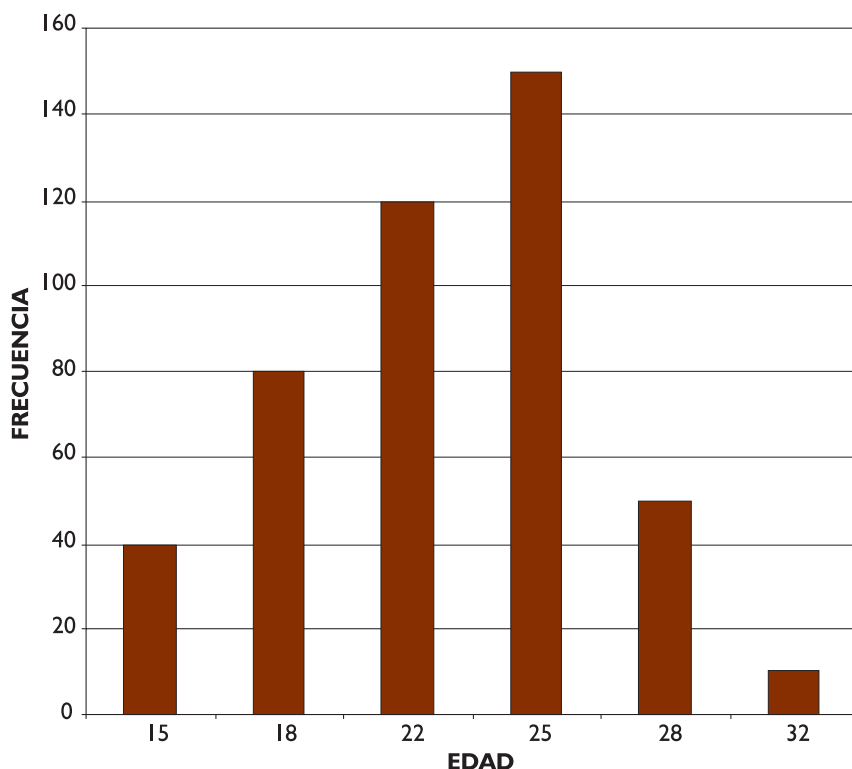
- **Polígono de frecuencias:** utilizando un eje de coordenadas, se colocan sobre el eje de abscisas los distintos valores de la variable y sobre el eje de ordenadas las frecuencias absolutas. A cada pareja de números formada por un valor de la variable y su frecuencia, le corresponde un punto en el plano, y luego se unen dichos puntos formando un polígono.

Ejemplo:

EDAD	FRECUENCIA
15	40
18	80
22	120
25	150
28	50
32	10



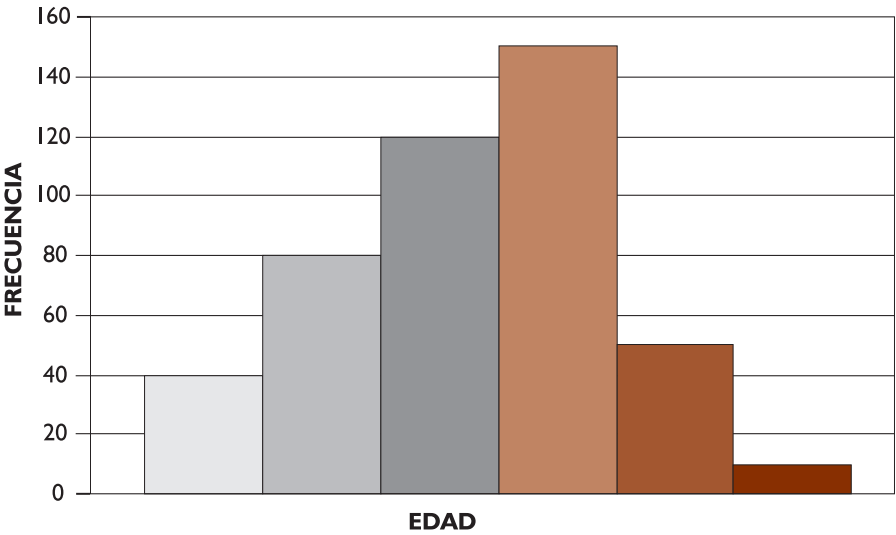
- **Diagrama de barras:** igual que en el gráfico anterior, se representa la variable en un eje de coordenadas y para hacer más visible el punto asociado a cada valor de la variable, se une con una línea desde el eje de abscisas. También puede hacerse horizontal, con solo representar en el eje de abscisas las frecuencias y en el de ordenadas los valores de la variable. También puede hacerse representando la variable en el eje de ordenadas y la frecuencia en el eje de abscisas. Con el mismo ejemplo anterior, un diagrama de barras quedaría como sigue:



- **Histograma:** se utiliza cuando la distribución es de tipo III, o sea, con intervalos. Si las clases tienen una amplitud constante se representa en un eje de coordenadas colocando en el eje de abscisas los diferentes límites de los intervalos y en el de ordenadas las frecuencias absolutas (o relativas). A cada intervalo le corresponderá una frecuencia con lo que se formará un rectángulo de altura determinada para cada intervalo. El área de dicho rectángulo será la frecuencia asociada a él. Cuando las clases tienen diferente amplitud no pueden representarse las frecuencias sino que recurriremos a la altura, simbolizada por h_1 , h_2 , ..., h_k que se calcula dividiendo la frecuencia absoluta de cada clase por su amplitud $h_i = n_i / a_i$ (siendo a_i la amplitud del intervalo en cuestión, que recordaremos que se calcula $L_i - L_{i-1}$)

Ejemplo:

EDAD	FRECUENCIA
15-18	40
18-22	80
22-25	120
25-28	150
28-32	50
32-35	10



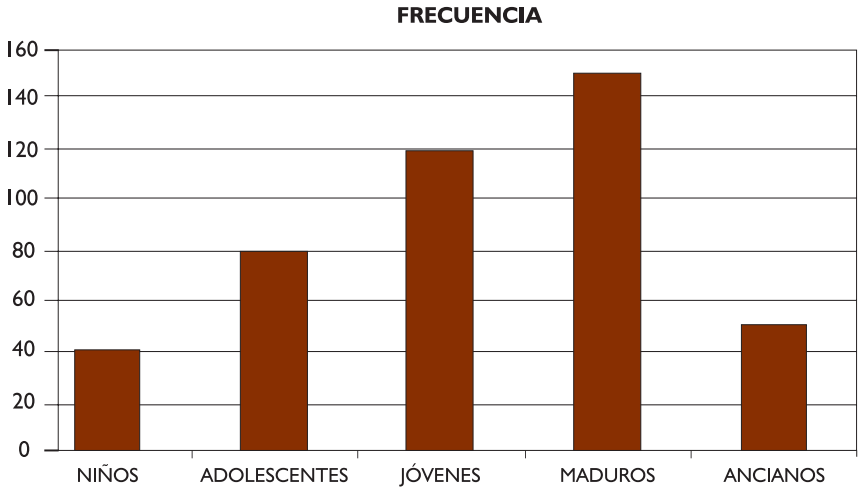
Hasta ahora hemos analizado características cuantitativas de las variables. En la estadística de atributos se analizan caracteres cualitativos, lo cual limita mucho las posibilidades de análisis, ya que muchos de los resultados que hemos estudiado anteriormente, aquí no tendrían sentido.

Las representaciones gráficas más usuales para caracteres cualitativos son:

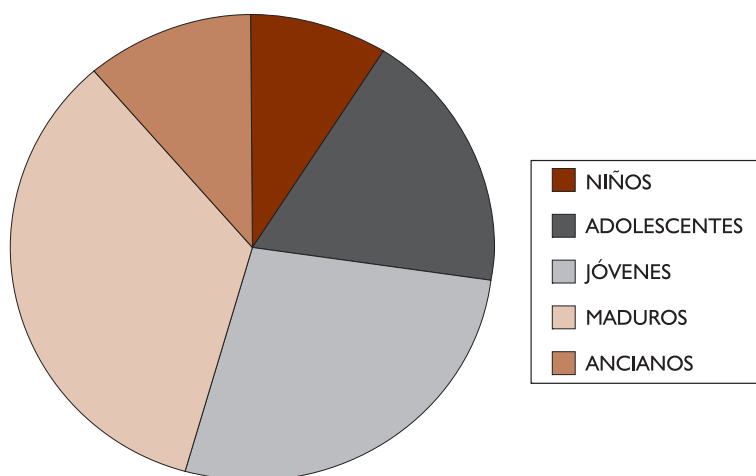
- **Diagrama de barras:** en un eje de coordenadas se presenta en las abscisas los caracteres cualitativos y en la ordenada las frecuencias absolutas. Sobre cada uno de los caracteres cualitativos de la abscisa se levantan rectángulos del mismo grosor y con la altura correspondiente a su frecuencia acumulada. Sería similar al analizado para variables cuantitativas pero en lugar de fijarse valores de la variable se especifica el carácter cualitativo de la misma.

Ejemplo:

EDAD	FRECUENCIA
NIÑOS	40
ADOLESCENTES	80
JÓVENES	120
MADUROS	150
ANCIANOS	50



- **Diagrama de sectores:** en un círculo se asigna un sector (porción) a cada uno de los caracteres cualitativos, siendo la amplitud del sector proporcional a la frecuencia del carácter. Esto se consigue haciendo corresponder 360° a la suma de todas las frecuencias (N) de los caracteres y hallando la correspondiente proporcionalidad (mediante reglas de 3). Obtendremos entonces los grados que en el círculo corresponde a cada carácter. Con el mismo ejemplo anterior, un diagrama de sectores quedaría como sigue:



- **Pictogramas:** cada carácter se representa por un dibujo de tamaño proporcional a la frecuencia del mismo. También es usual tomar un dibujo de tamaño estándar y repetirlo un número de veces proporcional a la frecuencia.
- **Cartogramas:** es la representación sobre mapas del carácter estudiado. Usualmente, las distintas modalidades que adopta este carácter se representan con colores distintos, rayados, punteos, etc.

6. Parámetros descriptivos

Los parámetros descriptivos son los que nos permiten obtener unas conclusiones acerca del fenómeno estudiado.

Pueden ser:

- Parámetros de tendencia central o promedios.
- Parámetros de dispersión.
- Parámetros de concentración.
- Parámetros de forma.

6.1. Parámetros de tendencia central o promedios

Al obtener de una población o de una muestra la distribución de frecuencias de una variable, lo que se persigue es reducir o condensar en pocas cifras el conjunto de observaciones relativas a dicha variable. Pero el proceso de

reducción hay que continuarlo hasta su grado máximo, es decir, hasta sustituir todos los valores observados por uno sólo, que se llama promedio.

La práctica estadística ha impuesto algunos promedios. En cada distribución se aplicarán aquel o aquellos que se consideren más representativos. Los más usuales son: media aritmética, mediana, moda, media cuadrática y media geométrica.

6.1.1. Media aritmética

La media aritmética es el número que se obtiene al dividir la suma de todas las observaciones por el número total de ellas. Se simboliza por $\bar{X}$ y en su cálculo se puede emplear la fórmula:

$$\bar{X} = \frac{\sum x_i n_i}{N}$$

Por tanto, la media aritmética es un valor de la variable, posiblemente no observable, que viene dado en la misma unidad de medida que la variable.

En el caso de distribuciones del tipo III, las x_i serán las marcas de clase.

Para distinguir la media de una población de la media de una muestra se emplean dos símbolos diferentes. Para la media de una población se utiliza el símbolo griego μ , mientras que el símbolo $\bar{X}$ se utiliza para la media de una muestra, aunque se calcula de idéntica forma. Esta distinción será importante cuando tratemos el tema de contraste de hipótesis. En este tema utilizaremos indistintamente una u otra notación.

6.1.2. Mediana

Si todos los valores de la variable se ordenan en sentido creciente, la mediana es el valor que ocupa en lugar central, o sea, el que deja a un lado y a otro el mismo número de observaciones. Se denota por Me. Por tanto, la mediana es el valor de la variable que corresponde a la primera frecuencia acumulada mayor que $N/2$.

En las distribuciones del tipo III se conocerá el intervalo en el que está contenida la mediana según el proceso anterior, pero no puede obtenerse exactamente la mediana. No obstante, puede realizarse una aproximación utilizando la siguiente fórmula:

$$Me = L_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i}$$

El hecho de que los intervalos de la distribución tengan o no una amplitud constante no afecta en absoluto para el cálculo de la mediana.

6.1.3. Moda

La moda es el valor de la variable que se presenta mayor número de veces, o sea, el valor más frecuente o con mayor frecuencia absoluta (n_i). La moda se simboliza por Mo .

En las distribuciones de tipo III no puede conocerse con exactitud la moda, aunque siguiendo la definición anterior, conoceremos en qué intervalo se encuentra y, con ello, se puede aproximar utilizando la fórmula:

$$Mo = L_{i-1} + a_i \frac{n_{i+1}}{n_{i-1} + n_{i+1}}$$

Esta fórmula presenta el inconveniente de no poder utilizarse cuando el intervalo modal sea el primero o el último. Para evitarlo, podemos utilizar esta otra fórmula que, aunque más complicada, no presenta ese inconveniente:

$$Mo = L_{i-1} + a_i \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

En el caso de que los intervalos no tengan la misma amplitud habrá que recurrir a las alturas, h_i . El intervalo modal será aquel que tenga mayor altura. Aplicaremos entonces la fórmula:

$$Mo = L_{i-1} + a_i \frac{h_{i+1}}{h_{i-1} + h_{i+1}}$$

Al igual que ocurriría en el caso anterior, para evitar el inconveniente que supone para su cálculo el hecho de que el intervalo modal sea el primero o el último, podemos recurrir a esta otra fórmula:

$$Mo = L_{i-1} + a_i \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})}$$

Si en la distribución existen dos modas se la denomina bimodal. Si existen más de dos modas, esta medida no será representativa de la distribución.

6.1.4. Media cuadrática

Se denota por C . Es útil en el caso de que la variable tome valores positivos y negativos, como en el caso del estudio de errores de medida. Puede interesarnos utilizarla como un promedio que no recoja los efectos del signo. La media cuadrática se calcula con la fórmula:

$$C = \sqrt{\sum X_i^2 \cdot n_i / N}$$

6.1.5. Media geométrica

Es poco usada en la práctica estadística por su complejidad de cálculo. Se denota por G y se calcula mediante la fórmula:

$$G = \sqrt[n]{\prod X_i^{n_i}}$$

6.1.6. Visión conjunta de los promedios

Ciñéndonos a los tres promedios más utilizados: media aritmética, mediana y moda, se cumple que,

$$\bar{X} \leq Me \leq Mo$$

O bien

$$Mo \leq Me \leq \bar{X}$$

En la mayoría de los casos los resultados obtenidos utilizando estos tres promedios difieren. Se nos plantea entonces la cuestión de qué promedio utilizar en cada caso.

La media aritmética viene influida por todos los valores de la variable, con lo que si existen valores muy pequeños o muy grandes de forma atípica con respecto al resto de las observaciones, la media aritmética se verá influenciada por ellos, cosa que no ocurre con la mediana y la moda.

Frente a este inconveniente, la media aritmética tiene la ventaja de utilizar toda la información recogida.

Por otra parte, la media aritmética tiene una formulación matemática, lo que permite su tratamiento científico, mientras que la mediana y la moda carecen de formulación.

La moda puede ser muy representativa en el caso de gran acumulación de variables en torno a un valor concreto.

6.2. Parámetros de posición: los cuantiles

Como hemos visto anteriormente, la mediana divide la distribución en dos partes iguales. Los cuantiles dividen la distribución en diversas partes iguales. Podemos clasificar los cuantiles en:

- **Cuantiles:** se expresan con Q_p y dividen la distribución en 4 partes iguales.
- **Deciles:** se expresan por D_p y dividen la distribución en 10 partes iguales.
- **Percentiles:** se expresan por P_p y dividen la distribución en 100 partes iguales.

Para calcularlos se utiliza el mismo método que en la mediana, pero no buscaremos el valor correspondiente a una frecuencia de $N/2$ sino del valor expresado por el cuantil.

De esta forma, para encontrar el cuantil calcularemos el valor de frecuencia acumulada siguiente:

Cuantiles: $Q_p = p.N/4$

Deciles: $D_p = p.N/10$

Percentiles: $P_p = p.N/100$

Así, el cuartil 1 es el primer valor observado que tenga una frecuencia absoluta acumulada superior a $1N/4$. El cuartil 2 es el valor que tenga una frecuencia absoluta acumulada de $2N/4$. El cuartil 3 es el valor cuya frecuencia absoluta acumulada supere a $3N/4$.

Lo mismo ocurrirá con los deciles. Por ejemplo, el decil 7 es aquel valor cuya frecuencia absoluta acumulada supere $7N/10$.

El percentil 65 será aquel valor cuya frecuencia absoluta acumulada supere $65N/100$.

Como puede suponerse, la mediana es, en realidad un cuantil. La mediana se corresponde con el cuartil 2, con el decil 5 y con el percentil 50.

En caso de distribuciones con intervalos, los cálculos serán iguales a los de la mediana, pero cambiando $N/2$ por el valor del cuantil.

$$Q_p = L_{i-1} + a_i \frac{PN/4 - N_{i-1}}{n_i}$$

$$D_p = L_{i-1} + a_i \frac{PN/10 - N_{i-1}}{n_i}$$

$$P_p = L_{i-1} + a_i \frac{PN/100 - N_{i-1}}{n_i}$$

6.3. Parámetros de dispersión

Nos ofrecen una medida de la dispersión de los valores en torno a un promedio. Los parámetros de dispersión más usuales son: recorrido, varianza, desviación estándar y coeficiente de variación. También pueden utilizarse la desviación media y la desviación mediana.

6.3.1. Recorrido o rango

Es la diferencia entre el mayor valor de la variable y el menor. También puede calcularse el cociente entre el mayor valor y el menor, que se conoce con el nombre de coeficiente de disparidad.

6.3.2. Varianza

Se denota por S^2 cuando se trata de la varianza de una muestra, mientras que emplearemos el símbolo griego σ^2 cuando se trate de la varianza de una población. Esta última es la suma del cuadrado de las diferencias entre cada valor de la variable y la media aritmética, dividido por el total de observaciones:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2 \cdot n_i}{N - 1}$$

También puede utilizarse esta otra fórmula (obteniendo el mismo resultado):

$$\sigma^2 = \frac{\sum X_i^2 \cdot n_i}{N} - (\mu)^2$$

La varianza de una muestra se calcula de forma similar, solo que dividiremos por $N - 1$, debido a los errores muestrales.

$$S^2 = \frac{\sum (X_i - \bar{X})^2 \cdot n_i}{N - 1}$$

La varianza tiene el inconveniente de que viene expresada en una unidad de medida que es el cuadrado de la unidad de medida de la variable. Para salvar este inconveniente se utiliza la desviación estándar.

6.3.3. Desviación estándar

También llamada desviación típica, es la raíz cuadrada positiva de la varianza. Se denota por S cuando se trata de la desviación típica de una muestra, mientras que la desviación típica de una población se indica σ . Viene expresada en la misma unidad de medida que la variable.

6.3.4. Coeficiente de variación de Karl Pearson

Se simboliza por V y es un número abstracto, con lo que podemos comparar la dispersión de diferentes distribuciones independientemente de la unidad de medida utilizada en cada una de ellas. Se calcula según la fórmula siguiente:

$$DM = \frac{\sum |X_i - M| \cdot n_i}{N}$$

Cuanto mayor sea la dispersión menos representativa será la media aritmética. Por tanto, si $V > 60\%$, la medida no es representativa y si $40\% < V < 60\%$, la media aritmética será poco representativa. Sólo cuando $V < 40\%$ tendremos una media aritmética representativa.

6.3.5. Desviación media

Se calcula según la fórmula:

$$DM = \frac{\sum |X_i - M| \cdot n_i}{N}$$

Viene expresada en la misma unidad de medida que la variable.

6.3.6. Desviación mediana

Se calcula según la fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N}$$

Estas dos últimas desviaciones permiten decidir si la mediana es o no más representativa que la media aritmética en la distribución. Bastaría con

comparar ambas y aquella que fuera menor supondría que la desviación producida por la medida de tendencia central es mejor. Por ello, si $DMe < DM$, diremos que la mediana es más representativa. Si se diera el caso contrario, sería más representativa la media aritmética.

6.4. Parámetros de concentración

Para estudiar la característica de la concentración de la variable se utiliza el índice de Gini.

$$G = 1 - \frac{\sum Q_i}{\sum P_i}$$

donde,

P_i = frecuencia relativa acumulada, multiplicada por 100.

Q_i = cada producto X_{ini} , haciéndolo relativo y acumulado, multiplicado por 100.

El índice de Gini oscila entre 0 y 1. Cuanto más próximo esté de 1, existirá mayor concentración de las observaciones en torno a un valor de la variable.

Podemos considerar que existe mucha concentración cuando el índice de Gini es superior a 0,75. En estos casos, la medida de tendencia central más representativa podría ser la moda, puesto que la moda indica el valor más habitual de la variable. Al estar la distribución muy concentrada, ese valor puede ser muy significativo de la medida de tendencia central.

6.5. Parámetros de forma

En este apartado vamos a analizar lo que se conoce como simetría de una distribución. Diremos que una variable tiene una distribución simétrica cuando en su histograma puede trazarse una línea vertical tal que al doblar por ella la figura, ambas partes coinciden exactamente. En caso contrario será asimétrica.

En las distribuciones campaniformes se usa el coeficiente de asimetría de Pearson, simbolizado por A, que se calcula con la siguiente fórmula:

$$A = \frac{\bar{X} - Mo}{S} \quad \text{o} \quad A = \frac{\mu - Mo}{\sigma}$$

- Si A es positiva, la variable tendrá una distribución asimétrica hacia la derecha.
- Si A es negativa, la variable tendrá una distribución asimétrica hacia la izquierda.
- Si A es cero, la variable tiene una distribución simétrica.

Podemos decir que cuando el coeficiente de asimetría es superior a 0,65 o bien inferior a $-0,65$ existe mucha asimetría (a derecha e izquierda, respectivamente). Cuando ello ocurre, suele ser habitual que la medida de tendencia central más significativa sea la Mediana, puesto que cuando existe mucha asimetría la media aritmética y la moda están muy distantes. Como recordaremos del estudio conjunto de los promedios, la media siempre se encuentra entre la media aritmética y la moda. Por ello, al ser una medida más suave puede resultar el promedio más representativo.

Ejercicios tema 7

1. En una encuesta realizada a 50 personas hemos encontrado los siguientes datos de estaturas en centímetros:

150, 185, 190, 172, 168, 171, 173, 175, 195, 186, 181, 176, 164, 168, 169, 181, 178, 175, 174, 171, 168, 165, 167, 174, 176, 179, 178, 182, 180, 186, 192, 174, 171, 170, 165, 169, 171, 172, 176, 174, 182, 195, 181, 174, 172, 176, 174, 168, 169, 170.

Confeccionar con ellos una tabla utilizando 5 clases de la misma amplitud y representar mediante un histograma estos datos.

2. Con los mismos datos del ejercicio anterior confeccionar una tabla con 4 clases de diferente amplitud: 10, 8, 8 y 19 respectivamente. Representar mediante un histograma dichos resultados.
3. Una encuesta realizada a 40 personas arroja la siguiente información acerca de sus ingresos anuales en euros:

10.000, 15.000, 14.000, 12.000, 14.000, 16.000, 25.000, 30.000, 32.000, 26.000, 22.000, 20.000, 18.000, 19.000, 17.000, 18.000, 16.000, 19.000, 18.000, 17.000, 12.000, 11.000, 18.000, 25.000, 22.000, 16.000, 19.000, 18.000, 17.000, 18.000, 19.000, 17.000, 35.000, 14.000, 19.000, 21.000, 18.000, 19.000, 17.000, 16.000.

Se pide: confeccionar una tabla con 5 clases de la misma amplitud. Calcular: media aritmética, mediana, moda, media cuadrática, varianza, desviación típica, coeficiente de variación de Pearson, desviación media, desviación mediana, índice de Gini y asimetría. Para simplificar los cálculos, los datos pueden expresarse en miles de euros.

4. En una encuesta sobre la edad en una muestra de personas que acuden un día determinado a un cine se ha obtenido el siguiente resultado:

5, 10, 25, 35, 45, 55, 65, 70, 5, 35, 70, 45, 35, 25, 25, 15, 25, 25, 35, 25, 45, 70, 65, 35, 25, 25, 55, 55, 15, 25, 45, 25, 35, 25, 25, 35, 25, 25, 15, 25, 35, 45, 25, 35, 55, 60, 25, 25, 15, 25.

Construir una tabla de tipo II y con ella determinar lo siguiente:

- Representación gráfica: diagrama de barras y polígono de frecuencias.
- ¿Qué porcentaje dentro del total de espectadores supone cada edad?

- Calcular la media aritmética.
- Calcular la mediana.
- Calcular la moda.

Construir una tabla de tipo III con 7 intervalos regulares y con ella determinar lo siguiente:

- Representación gráfica: histograma.
- Calcular la media aritmética.
- Calcular la mediana.
- Calcular la moda.

- 5.** En una encuesta sobre el número de veces que una muestra de consumidores acude a un cierto establecimiento durante un año se ha obtenido el siguiente resultado:

5, 10, 25, 5, 45, 5, 6, 10, 5, 5, 50, 5, 35, 5, 25, 5, 25, 5, 5, 25, 5, 8, 5, 35, 25, 0, 5, 15, 5, 25, 5, 25, 5, 8, 25, 5, 5, 25, 7, 5, 35, 5, 6, 35, 5, 8, 5, 25, 15, 7.

Construir una tabla de tipo III con 5 intervalos regulares y con ella determinar lo siguiente:

- Representación gráfica: histograma.
- Calcular la media aritmética.
- Calcular la mediana.
- Calcular la moda.

Construir una tabla de tipo III con 4 intervalos de diferente amplitud (15, 10, 10 y 15) y con ella determinar lo siguiente:

- Representación gráfica: histograma.
- ¿Qué porcentaje dentro del total supone cada intervalo?
- Calcular la media aritmética.
- Calcular la mediana.
- Calcular la moda.

- 6.** En una encuesta realizada a las 500 familias de una urbanización acerca del número de coches de su familia se ha obtenido el siguiente resultado:

Nº de coches	Nº encuestados
0	20
1	180
2	220
3	50
4	30

- Representar gráficamente la distribución de dos formas diferentes.
 - ¿Cuál es el número de coches más habitual en las familias?
 - ¿Es representativa la media aritmética?
 - ¿Sería más representativa la mediana?
 - ¿Cuál es el porcentaje de familias con 2 coches?
7. En una encuesta sobre el consumo de cerveza mensual en las familias se ha encuestado a 1.300 personas obteniendo el siguiente resultado:

Latas de cerveza al mes	Nº de encuestados
Entre 0 y menos de 20	80
Entre 20 y menos de 30	180
Entre 30 y menos de 40	350
Entre 40 y menos de 60	480
Entre 60 y menos de 90	150
Entre 90 y 130	60

- Representar gráficamente la distribución.
- ¿Cuál es el consumo de latas de cerveza al mes más habitual en las familias?
- ¿Es representativa la media aritmética?
- ¿Sería más representativa la mediana?
- ¿Cuál es el porcentaje de familias con consumo de cerveza en cada tramo?

8. En una encuesta realizada a una muestra de 900 personas acerca del número de componentes de la unidad familiar se ha obtenido el siguiente resultado:

Nº de componentes	Nº encuestados
1	320
2	280
3	220
4	50
5	30

- Representar gráficamente la distribución de dos formas diferentes.
 - ¿Cuál es el número de componentes más habitual en las familias encuestadas?
 - ¿Es representativa la media aritmética?
 - ¿Sería más representativa la mediana?
9. En una encuesta a la población de un municipio de 1.050 hogares sobre el consumo eléctrico mensual en € se ha obteniendo el siguiente resultado:

Consumo eléctrico €	Nº de encuestados
Entre 0 y menos de 50	480
Entre 50 y menos de 100	280
Entre 100 y menos de 150	150
Entre 150 y menos de 200	80
Entre 200 y menos de 250	50
Entre 250 y 300	10

- Representar gráficamente la distribución.
- ¿Cuál es el consumo eléctrico al mes más habitual?
- ¿Es representativa la media aritmética?
- ¿Sería más representativa la mediana?

- 10.** En un estudio sobre los gustos de los consumidores acerca de un preparado lácteo se ha obtenido la siguiente información:

Sabor	Nº encuestados
Fresa	60
Natural	180
Limón	120
Coco	140
Piña	100

- ¿Qué medida de tendencia central podemos obtener y cuál es su resultado?
 - ¿Qué porcentaje de encuestados prefieren cada sabor?
 - Representa esa distribución con un diagrama de barras.
- 11.** En una encuesta sobre el número de viviendas en propiedad que posee la unidad familiar se ha obtenido las respuestas siguientes:
- 1, 1, 2, 1, 3, 2, 1, 4, 5, 4, 1, 2, 1, 3, 1, 2, 1, 1, 2, 1, 3, 1, 2, 1, 1, 1, 2, 1, 2, 1.
- Confeccionar una tabla de tipo II.
 - ¿Sería más representativa la media aritmética o la mediana?
 - Representa esa distribución con un polígono de frecuencias.
- 12.** Un estudio acerca del gasto semanal de compra en grandes superficies realizado a 200 personas se ha obtenido el siguiente resultado:

Consumo semanal en €	Nº encuestados
Más de 0 y menos de 50	10
De 50 a menos de 80	40
De 80 a menos de 100	80
De 100 a menos de 130	30
De 130 a menos de 150	20
De 150 a 200	20

- ¿Es representativa la media aritmética?
- Representar gráficamente esta distribución.

- 13.** En una encuesta sobre los hábitos de lectura se ha preguntado a los entrevistados sobre el número de libros que leen cada año y se ha obtenido el siguiente resultado:

0, 1, 2, 3, 4, 12, 15, 25, 30, 50, 10, 5, 4, 1, 0, 0, 6, 12, 18, 42, 20, 23, 4, 16, 15, 26, 11, 28, 35, 36, 17, 5, 7, 9, 3, 2, 0, 1, 14, 4.

- Construir una tabla de amplitud regular con 5 intervalos.
 - En base a su concentración ¿sería más representativa la moda que cualquier otra medida de tendencia central?
 - Calcular la moda y comentar este resultado.
- 14.** En una encuesta acerca del consumo de combustible en euros semanal se ha obtenido el siguiente resultado:

Consumo en €	Nº personas
(0-10)	80
(10-30)	90
(30-40)	40
(40-60)	20
(60-100)	10

- Según su asimetría ¿podemos decir que la mediana es la tendencia central más representativa de esta distribución? Explica tu respuesta y su significado.
 - Calcular la mediana y comenta este resultado.
- 15.** En una encuesta sobre el número de veces que los vecinos de una urbanización van al cine durante un año se ha obtenido el siguiente resultado:
- 0, 1, 2, 3, 4, 12, 15, 25, 30, 50, 10, 5, 4, 1, 0, 0, 6, 12, 18, 42, 20, 23, 4, 16, 15, 26, 11, 28, 35, 36, 17, 5, 7, 9, 3, 2, 0, 1, 14, 4.
- Construir una tabla de tipo III con 4 intervalos irregulares de amplitud 10, 10, 15 y 15, respectivamente.
 - Realiza un estudio comentado y completo sobre la medida de tendencia central que pueda ser más representativa.

Tema 8

La estadística bivalente

1. Tablas de presentación de datos bivariantes

En los análisis estadísticos es muy frecuente que el estudio de un fenómeno suponga la existencia de una relación entre dos variables. Tendremos ahora pares de valores, X e Y , asociados.

Podemos considerar 3 tipos de presentación de los datos:

TIPO I: Es el caso de pocas observaciones, con lo cual habrá pocos pares de datos. Se disponen los pares de valores en dos columnas. Ejemplo: análisis de horas totales de estudio con calificaciones medias obtenidas en los exámenes en 5 alumnos:

Horas : X	Calificaciones: Y
12	4
20	6
23	7
28	8
35	9

TIPO II: Cuando hay muchas observaciones, pero pocos pares distintos de datos. Se presentan mediante 3 columnas, las dos primeras para las variables y la tercera para las frecuencias, n_i . Ejemplo: análisis en 130 familias según el número de hijos con el gasto anual en viajes turísticos:

Hijos: X	Gasto: Y	Frecuencia: n_i
0	2.000,00	20
0	3.000,00	10
1	2.000,00	40

Hijos: X	Gasto: Y	Frecuencia: n_i
1	3.000,00	20
1	4.000,00	10
2	1.500,00	20
2	2.500,00	5
2	4.000,00	1
3	2.000,00	3
3	4.500,00	1

TIPO III: cuando el número de observaciones es muy elevado y el número de pares de valores también. Se disponen, en este caso, en lo que se llama tabla de doble entrada. Una variable en una columna y la otra en una fila, de forma que en la intersección de un par de valores, fila y columna, aparezcan las frecuencias. Ejemplo: análisis en 75 países del número de turistas recibidos durante 2003 con los ingresos por turismo en ese año:

I N G R E S O S	11.000	0	0	0	1	3	2
	10.000	0	0	2	4	7	1
	9.000	0	1	4	8	3	1
	8.000	1	1	5	7	2	0
	7.000	1	2	6	5	0	0
	6.000	5	2	1	0	0	0
		10	20	30	40	50	60

TURISTAS EN MILLONES

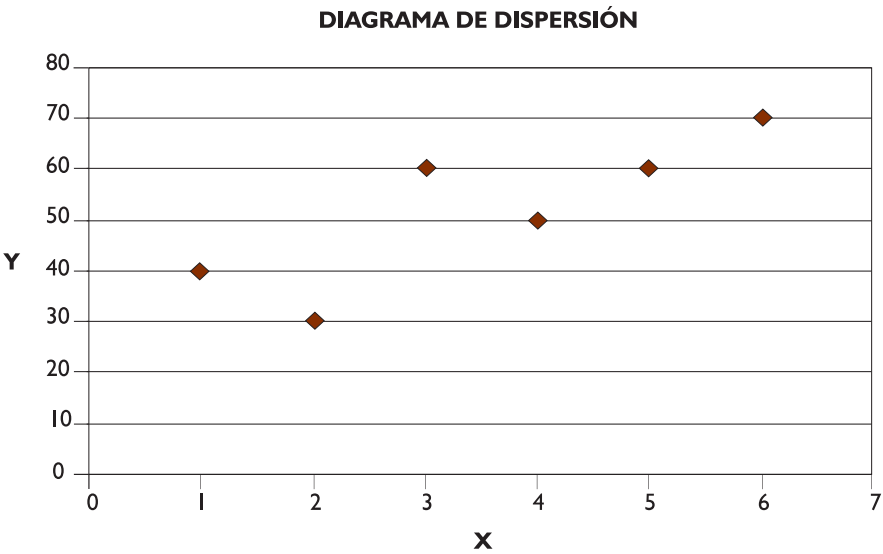
Con estas distribuciones se pretende conocer la característica de relación o dependencia que exista entre las dos variables.

2. Diagrama de dispersión

En eje de coordenadas se colocan los valores de X en el eje de abscisas y los de Y en el eje de ordenadas. Cada par de valores X e Y supondrá un punto dentro del gráfico. Se obtiene así una nube de puntos, llamado diagrama de dispersión. Este diagrama es un claro indicador de la dependencia existente entre las dos variables.

Ejemplo: Dada la siguiente distribución de valores de X e Y, confeccionamos el diagrama de dispersión.

X	Y
1	40
2	30
3	60
4	50
5	60
6	70



3. Regresión

El análisis de la regresión consiste en obtener la línea “ideal”, denominada línea de regresión, hacia la cual tienden los puntos de un diagrama de dispersión.

Esta línea de regresión puede obtenerse de muy diversas formas, pero el método más utilizado es el analítico, que usa una función matemática para explicar la dependencia de una variable con respecto a la otra. Si nuestras variables son X e Y, considerando que la variable Y depende de la variable X, podemos escribir:

$$Y^* = f(x)$$

Donde Y^* serán los valores de Y obtenidos al aplicar una función matemática a los diferentes valores de X .

Habrà, pues, una diferencia entre los valores reales de Y y los obtenidos como Y^* , que serán los errores de nuestra línea de regresión:

$$Y - Y^* = e$$

El problema que nos surge ahora es la selección de la función matemática $f(x)$.

Las funciones más utilizadas son:

Recta o función lineal:

$$Y^* = a + b \cdot X$$

Parábola de segundo grado:

$$Y^* = a + b \cdot X + c \cdot X^2$$

Función potencial:

$$Y^* = a \cdot X^b$$

Función exponencial:

$$Y^* = a \cdot b^X$$

El problema que surge de inmediato es calcular el valor exacto de los parámetros a , b y c .

El método más utilizado es el de los mínimos cuadrados.

4. Ajuste por mínimos cuadrados

Vamos a aplicar este método únicamente para el caso de una función lineal (recta):

$$Y^* = a + b \cdot X$$

Este método proporciona unos valores numéricos para los parámetros a y b con la condición:

$$\sum e_i^2 = \text{mínimo}$$

O sea, que la suma de los errores al cuadrado sea mínima, lo cual puede escribirse:

$$\sum \sum (Y_j - Y_{ij}^*)^2 n_{ij} / N = \text{mínima}$$

Como hemos elegido de modelo una recta, tendremos:

$$\sum \sum (Y_j - (a + b \cdot X_i))^2 n_{ij} / N = \text{mínima}$$

Para conseguir un mínimo habremos de igualar a 0 la primera derivada con respecto al parámetro “a” y la primera derivada con respecto a “b”.

De esta forma obtendremos un sistema de dos ecuaciones con dos incógnitas. Una vez resuelto tendremos:

$$a = \mu_y - b\mu_x$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

Donde σ_{xy} es lo que se conoce como **covarianza**, definida por la fórmula

$$\sigma_{xy} = \frac{\sum \sum (X_i - \mu_x)(Y_j - \mu_y)n_{ij}}{N}$$

O esta otra, quizás más fácil de calcular:

$$\sigma_{xy} = \frac{\sum \sum X_i Y_j n_{ij}}{N} - \mu_x \cdot \mu_y$$

De esta forma, conseguir la ecuación de la recta es fácil, ya que si tenemos los valores de “a” y de “b”, nuestra ecuación será:

$$Y^* = a + b \cdot X$$

Por tanto, para predecir el comportamiento de Y, en función de los valores de X, sólo habrá que sustituir X por el valor a analizar y obtendremos el correspondiente a Y.

5. Varianza residual y coeficiente de determinación

Toda ecuación de regresión debe venir acompañada de la medida de su representatividad o bondad. En consecuencia, una medida de la dispersión de las Y_j observadas con respecto a las Y^*_i calculadas según nuestra ecuación es un poderoso e imprescindible instrumento complementario en el análisis de la regresión.

La medida de dispersión más utilizada es la varianza residual, simbolizada por S_e^2 , que se calcula con la fórmula:

$$S_e^2 = \frac{\sum \sum (Y_j - Y_i^*)^2 \cdot n_{ij}}{N}$$

Para poder realizar comparaciones entre la bondad de diversos ajustes se utiliza el coeficiente de determinación, simbolizado por R^2 , que se calcula con la fórmula:

$$R^2 = 1 - \frac{S_e^2}{\sigma_y^2}$$

Toma valores entre 0 y 1. Cuanto más próximo esté al 1, mayor bondad tendrá el ajuste realizado. Cuando $R^2 < 0,75$ habrá que pensar en utilizar otro modelo de ecuación que se ajuste mejor a los datos observados. Una situación muy satisfactoria será cuando $R^2 > 0,90$.

Una utilidad importante del ajuste mínimo cuadrático es la posibilidad de predicción. Una vez elegido el modelo y comprobada su bondad, podemos hacer consideraciones del valor de la variable Y cuando X tenga un valor teórico, no observado. Pero esa predicción será más deficiente cuanto más nos alejemos del recorrido observado de la variable X.

6. Correlación lineal simple

La correlación busca un número, denominado coeficiente de correlación, para indicar objetivamente el grado de variación conjunta que tienen las variables.

Para el caso de dos variables se usa el coeficiente de variación lineal, simbolizado por “r”, que se calcula con la fórmula:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Este coeficiente toma valores entre -1 y $+1$. Será positivo cuando las variables se muevan en el mismo sentido, es decir, cuando al aumentar una aumente la otra. Es negativo cuando las variables se mueven en sentido

contrario. Cuanto más cerca esté de los extremos (-1 o $+1$) mayor correlación habrá entre las variables.

En general, podemos decir que cuando este coeficiente toma valores superiores a $0,85$ existe una relación lineal directa entre las variables. Cuando tome valores inferiores a $-0,85$ existirá una relación lineal inversa. Con valores entre $-0,85$ y $0,85$ diremos que puede no existir relación o, si existe, ésta no es lineal.

La correlación señala el grado de dependencia entre dos variables sin exigir ningún tipo de relación especial entre ellas, es decir, no nos informa acerca de cual es la variable dependiente y cual es la variable explicativa.

En el caso de regresión lineal, se da que $r^2 \approx R^2$.

Ejercicios tema 8

1. La siguiente tabla muestra un estudio acerca del consumo eléctrico asociado a la ocupación de un hotel. Se han analizado los últimos 60 meses.

% Ocupación	Consumo KW	Frecuencia
40-60	1.000	5
60-70	1.200	15
70-80	1.500	25
80-90	1.700	10
90-100	2.000	5

Se pide: confeccionar un diagrama de dispersión, calcular la recta de regresión y el coeficiente de determinación. Calcular el coeficiente de correlación lineal.

2. La siguiente tabla muestra un estudio acerca del número de días de vacaciones de una serie de 1.000 personas asociado a su nivel de renta en miles de euros.

Renta	días	Frecuencia
10-15	6	50
10-15	7	120
15-20	7	180
15-20	9	220
20-25	8	150
20-25	12	140
25-30	14	100
30-35	15	40

Se pide: confeccionar un diagrama de dispersión, calcular la recta de regresión y el coeficiente de determinación. Calcular el coeficiente de correlación lineal. ¿Podemos decir, según este estudio, que el número de días de vacaciones depende del nivel de renta?

3. La siguiente tabla muestra un estudio de mercado acerca del consumo mensual de cierto producto en función del número de personas que compo-

nen la familia. Se ha realizado una encuesta a 800 personas, obteniendo el siguiente resultado:

Nº familiares	Consumo	Encuestados
1	30	100
2	50	200
3	60	300
4	80	150
5	100	50

Se pide: confeccionar un diagrama de dispersión, calcular la recta de regresión y el coeficiente de determinación. Calcular el coeficiente de correlación lineal ¿Cuál sería el consumo de una familia de 6 miembros si la relación entre las variables es lineal?

4. Se ha realizado una encuesta a 1.000 personas acerca del número de veces que acuden a un cine en el año, en función de la edad del encuestado, obteniendo los resultados que se presentan en la tabla siguiente:

Edad	Asistencia al cine	Encuestados
0-10	40	50
10-20	50	100
20-30	40	150
30-40	35	200
40-50	25	200
50-60	20	100
60-70	10	100
70-80	5	100

Se pide: confeccionar un diagrama de dispersión, calcular la recta de regresión y el coeficiente de determinación. Calcular el coeficiente de correlación lineal. ¿Podemos decir que el número de veces que se asiste al cine depende de la edad, según este estudio?

5. La siguiente tabla muestra un estudio acerca de los kilómetros anuales que se realizan en coche en función del sueldo mensual de la persona, encuestando a 900 personas.

Sueldo	Km anuales	Encuestados
500-1000	20.000	150
1000-1500	30.000	200
1500-2000	15.000	250
2000-2500	35.000	100
2500-3000	20.000	200

Se pide: confeccionar un diagrama de dispersión, calcular la recta de regresión y el coeficiente de determinación. Calcular el coeficiente de correlación lineal ¿Podemos decir que existe relación lineal entre el sueldo mensual y los kilómetros mensuales que se realizan en vehículo, en función de este estudio?

6. Se ha analizado la titulación académica de 1.100 personas clasificándola de 1 a 8 (siendo 8 la de doctorado y 1 sin estudios) para comprobar si existe relación entre ella y el número de países visitados. Todos los encuestados tenían una edad que oscilaba entre 35 y 45 años. Los resultados se muestran en la tabla siguiente:

Estudios	Países visitados	Encuestados
1	12	150
2	4	120
3	5	180
4	10	200
5	8	150
6	10	150
7	12	100
8	5	50

Se pide: confeccionar un diagrama de dispersión, calcular la recta de regresión y el coeficiente de determinación. Calcular el coeficiente de correlación lineal. ¿Podemos decir que el número de países visitados depende de la titulación académica de forma lineal, según este estudio?

Tema 9

Las series temporales

1. Introducción

Una serie temporal está constituida por un conjunto de valores de una variable relacionado con otro conjunto de instantes o intervalos de tiempo. Como el tiempo, a su vez, es otra variable, también podemos decir que una serie temporal no es más que la consideración conjunta de dos variables, siendo el tiempo la variable independiente o explicativa y la otra variable es la dependiente.

Como el tiempo es una variable que implica un orden natural, las observaciones se escriben siguiendo precisamente este orden, de aquí la denominación de “serie”, que supone una sucesión de términos en un cierto orden. Además de serie temporal también se le llama serie cronológica o histórica.

El análisis de una serie temporal permite, por un lado, describir la evolución pasada de una variable y, por otro, formular predicciones sobre un futuro más o menos cercano.

2. Análisis de una serie temporal

La variable no temporal de una serie cronológica es una función del tiempo. Sea Y la variable y “ t ” el tiempo, entonces:

$$Y = f(t)$$

Las variaciones temporales de la variable son el resultado de la conjunción de las cuatro fuerzas o componentes siguientes:

1. **Tendencia secular** o simplemente **tendencia**: es la que refleja la evolución a largo plazo de la serie histórica. La tendencia expresa si la serie es estacionaria o evolutiva. Al considerar estos movimientos a largo plazo se ignoran conscientemente las variaciones de la variable a medio y corto plazo.

- 2. Fluctuaciones cíclicas:** son movimientos a plazo medio que se repiten casi de forma periódica, pero no son tan regulares como la tendencia. A veces, resulta difícil separar ambos componentes. En estos casos, es útil englobar ambos en uno solo, denominados ciclo-tendencia o tendencia generalizada.
- 3. Variaciones estacionales:** son movimientos a corto plazo que se repiten con carácter periódico. La repetición de estos movimientos se debe a las estaciones del año o a los hábitos o costumbres de la población. El período de repetición puede ser el año, las estaciones, los meses, las semanas, los días e incluso divisiones del día.
- 4. Variaciones accidentales,** también llamado movimiento irregular: son las provocadas por factores fortuitos estrictamente aleatorios o por factores esporádicos con repercusiones sensibles en la serie temporal, tales como inundaciones, incendios, guerras, accidentes, etc.

Los dos esquemas más admitidos sobre la forma en que una serie histórica se descompone en los cuatro componentes considerados son el aditivo y el multiplicativo.

Sea,

T= Tendencia

C= Fluctuaciones cíclicas

E= Variaciones estacionales

I= Movimiento irregular

El esquema aditivo supone que:

$$Y = T + C + E + I$$

Y el esquema multiplicativo considera que:

$$Y = T * C * E * I$$

La razón de utilizar una de estas dos hipótesis radica fundamentalmente en la sencillez y operatividad de las mismas.

Nos surge ahora el problema de determinar aquel esquema que mejor se adapte a nuestras observaciones. Existen diversos métodos, pero quizás el más empleado por su rapidez y facilidad es el análisis del gráfico de desviaciones estándar-medias.

Para ello, se calculan las medias aritméticas y desviaciones típicas para los datos de cada año. Luego se representan en un eje de coordenadas, colocando en el eje de abscisas las medias y en el eje de ordenadas las desviaciones estándar. Cada par de valores (media-desviaciones estándar) supondrá un punto en el gráfico.

Si los puntos representados están situados en una paralela al eje de abscisas se aplicará el esquema aditivo. Si están situados, aproximadamente, sobre una recta que forma un cierto ángulo con el eje de abscisas, se aplicará el esquema multiplicativo.

3. Cálculo de la tendencia

Existen muchos métodos para estudiar u obtener la tendencia. Los más empleados son: medias móviles y ajuste de una función. En el epígrafe siguiente veremos el método de las medias móviles. Ahora veremos el método de ajuste de una función.

Se trata de obtener una función matemática que exprese la tendencia. Una de las variables es el tiempo (que será la variable explicativa, X). La otra variable (dependiente) serán las medias anuales (Y).

El problema que nos surge ahora es la selección de la función matemática $f(x)$. Las funciones más utilizadas son:

Recta o función lineal:

$$Y^* = a + b \cdot X$$

Parábola de segundo grado:

$$Y^* = a + b \cdot X + c \cdot X^2$$

Función potencial:

$$Y^* = a \cdot X^b$$

Función exponencial:

$$Y^* = a \cdot b^X$$

Ahora es necesario calcular el valor exacto de los parámetros a , b y c . El método más utilizado es el de los mínimos cuadrados, como vimos en el tema 8.

Como recordaremos, los valores de “a” y “b” en una recta son los siguientes:

$$a = \mu_y - b\mu_x$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

Ejemplo:

La tabla siguiente muestra el porcentaje de ocupación hotelera en una zona turística en los últimos 5 años, estudiada por trimestres.

Años	Trimestres			
	1º	2º	3º	4º
2004	40	60	55	45
2005	50	65	60	50
2006	55	65	65	55
2007	60	75	70	55
2008	70	80	75	65

Solución:

Calcularemos las medias anuales de ocupación:

$$2004 = (40+60+55+45)/4 = 50$$

$$2005 = (50+65+60+50)/4 = 56,25$$

$$2006 = (55+65+65+55)/4 = 60$$

$$2007 = (60+75+70+55)/4 = 65$$

$$2008 = (70+80+75+65)/4 = 72,5$$

Con estos datos confeccionamos una tabla donde calcularemos el resto de la información necesaria para ajustar una función.

X	Y	$X_i * Y_j$	X^2
2004	50	100200	4016016
2005	56,25	112781,25	4020025
2006	60	120360	4024036
2007	65	130455	4028049
2008	72,5	145580	4032064
10030	303,75	609376,25	20120190

$$\text{Media de } X = 10030/5 = 2006$$

$$\text{Media de } Y = 303,75 / 5 = 60,75$$

$$\text{Varianza de } X (\sigma_x^2) = 20120190 / 5 - (2006)^2 = 2$$

$$\text{Covarianza } (\sigma_{xy}) = 609376,25 / 5 - (2006 * 60,75) = 10,75$$

Por tanto,

$$b = 10,75 / 2 = 5,375$$

$$a = 60,75 - (5,375) * 2006 = -10.721,5$$

Así tenemos que la ecuación de la recta será:

$$Y = -10.721,5 + 5,375X$$

4. Medias móviles

Este es un método mecánico para suavizar las irregularidades y fluctuaciones de la serie temporal y obtener así la línea de tendencia. Consiste en tomar un cierto período de tiempo que comprende varias observaciones consecutivas de las cuales se calcula su media aritmética. El período es constante y se va haciendo desplazar a lo largo de la serie. Cada media obtenida se hace corresponder al tiempo central del período.

Normalmente, se elige como período para calcular las medias móviles el año. Si, por ejemplo, en la serie se ha estudiado la variable por trimestres, la media móvil se hará para cuatro observaciones. La siguiente media móvil se hará suprimiendo la primera observación y añadiendo la quinta. Y así, sucesivamente. Pero el período considerado podría ser superior o inferior al año. Hay que tener en cuenta que cuanto mayor sea el período más se suavizará la serie pero también se pierde información.

Si dentro del período elegido el número de observaciones es impar, el resultado de la media móvil corresponderá con un momento de la serie. Se dice entonces que la media móvil está centrada.

Si, por el contrario, el número de observaciones que comprende el período es par, las medias móviles estarán descentradas ya que no coinciden con un momento concreto de la serie, sino con un momento situado entre dos observaciones. Para centrar estas medias volverá a utilizarse este método considerando como período el que corresponde a dos medias móviles descentradas, obteniendo así medias móviles centradas.

Ejemplo:

Utilizando el mismo ejemplo anterior:

	Trimestres			
Años	1°	2°	3°	4°
2004	40	60	55	45
2005	50	65	60	50
2006	55	65	65	55
2007	60	75	70	55
2008	70	80	75	65

El cálculo de las medias móviles se haría tomando los períodos de 4 en 4. De esta forma, resultaría:

50	61,25
52,5	63,75
53,75	65
55	65
56,25	67,5
57,5	68,75
57,5	70
58,75	72,5
60	

Las medias móviles no estarían centradas, puesto que coincidirían con un momento entre dos períodos. Por ello, volvemos a calcular las medias móviles, tomadas ahora de 2 en 2. Esos resultados los llevamos a una nueva tabla:

	Trimestres			
Años	1°	2°	3°	4°
2004			51,2500	53,1250
2005	54,3750	55,6250	56,8750	57,5000
2006	58,1250	59,3750	60,6250	62,5000
2007	64,3750	65,0000	66,2500	68,1250
2008	69,3750	71,2500		

5. Variación estacional

El cálculo de las variaciones estacionales dependerá del esquema estimado, aditivo o multiplicativo. Lo que se pretende determinar es la influencia de la estacionalidad en la serie.

5.1. Esquema multiplicativo

El método más empleado en este esquema es el de la razón a la media móvil. Se calcula la tendencia por el método de las medias móviles. Una vez centradas, se habrá suprimido el componente estacional e irregular.

Como la serie temporal es:

$$Y = T * C * E * I$$

Y la media móvil centrada es:

$$\text{Media móvil} = T * C$$

Si dividimos cada observación de una serie original por las medias móviles obtenidas, conseguiremos aislar el componente estacional y el irregular.

$$\frac{T * C * E * I}{T * C} = E * I$$

Si el componente irregular no es muy significativo, estos índices pueden ser estables para cada estación. En este caso, los índices generales de variación estacional pueden obtenerse tomando un promedio de todos los índices para ese subperíodo o estación. Pero si la estacionalidad es cambiante o evolutiva, tal resumen no es aconsejable.

Ahora bien, si existe estacionalidad y hemos hallado los índices generales de variación estacional mediante un promedio, estos índices no deben afectar al nivel de la serie, por lo que es razonable exigir que la media de los índices generales de variación sea 1, o lo que es lo mismo, su suma sea igual al número de subperíodos en los que se divide el año.

Si no es así, podemos dividir el resultado de su suma por el número de subperíodos del año y esta cantidad dividirá a cada índice general de variación estacional. Una vez hecho esto, se dice que los índices de variación estacional están normalizados.

Ejemplo:

Siguiendo el mismo ejemplo anterior, tomando ya la tabla de medias móviles calculadas, en el esquema multiplicativo hemos de dividir la serie original por los valores hallados de medias móviles. Los resultados se expresan en la siguiente tabla:

Años	Trimestres			
	1º	2º	3º	4º
2004			1,073171	0,847059
2005	0,919540	1,168539	1,054945	0,869565
2006	0,946237	1,094737	1,072165	0,880000
2007	0,932039	1,153846	1,056604	0,807339
2008	1,009009	1,122807		

Calculamos ahora un promedio con estos índices, obteniendo para cada estación (trimestre en este ejemplo) un valor de variación estacional:

$$1^\circ = 0,951706$$

$$2^\circ = 1,134982$$

$$3^\circ = 1,064221$$

$$4^\circ = 0,850991$$

Para que los índices estén normalizados es necesario que su suma sea igual a 4. En este caso, la suma es 4,001900. Para que el resultado sea 4, hemos de dividir cada uno de los índices hallados por el resultado anterior dividido por 4, o sea 1,000475123. Así obtenemos:

$$1^\circ = 0,951254$$

$$2^\circ = 1,134443$$

$$3^\circ = 1,063716$$

$$4^\circ = 0,850587$$

Estos serán los índices de variación estacional normalizados.

5.2. Esquema aditivo

Suponemos ahora que

$$Y = T + C + E + I$$

Por tanto, una vez calculada la tendencia, restaremos a la serie original la serie de tendencias, con lo que habremos conseguido aislar el componente estacional.

En el caso de estacionalidad estable, se calculan las medias para cada estación. Para que se cumpla que la variación estacional no afecte a los niveles de la serie, la suma de esas medias debe ser igual a cero. En caso contrario dividiremos el resultado de su suma por el número de subperíodos en los que se divide el año y restaremos esa cantidad (que puede ser positiva o negativa, por lo que si fuera negativa, se sumaría) a cada valor de estacionalidad antes obtenido. El resultado nos indicará la influencia de la estación en la serie, según el esquema aditivo.

Ejemplo:

Siguiendo el ejemplo anterior, en el esquema aditivo restaríamos a la serie original las medias móviles halladas. Los resultados aparecen en la tabla siguiente:

Años	Trimestres			
	1º	2º	3º	4º
2004			3,750000	-8,125000
2005	-4,375000	9,375000	3,125000	-7,500000
2006	-3,125000	5,625000	4,375000	-7,500000
2007	-4,375000	10,000000	3,750000	-13,125000
2008	0,625000	8,750000		

Calculamos las medias para cada estación (trimestre)

$$1^\circ = -2,8125$$

$$2^\circ = 8,4375$$

$$3^\circ = 3,75$$

$$4^\circ = -9,0625$$

Sumamos estas medias para comprobar si el resultado es igual a 0. Vemos que su suma es igual a 0,3125. Dividimos este resultado entre los 4 períodos del año, o sea 0,078125. Restamos esta cantidad a cada índice, quedando los siguientes resultados:

$$1^{\circ} = -2,890625$$

$$2^{\circ} = 8,359375$$

$$3^{\circ} = 3,671875$$

$$4^{\circ} = -9,140625$$

que serán los índices de variación estacionales según el esquema aditivo.

Ejercicios tema 9

1. La tabla siguiente muestra el porcentaje de ocupación hotelera en una zona turística durante 5 años, estudiada por cuatrimestres.

Años	Cuatrimestres		
	1º	2º	3º
2007	40	60	55
2008	50	65	60
2009	55	65	65
2010	60	75	70
2011	70	80	75

- Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema multiplicativo.
- Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema aditivo.
- Determinar cuál de los dos esquemas es el más apropiado para este caso.
- Calcular una recta (por mínimos cuadrados) que explique la tendencia según las medias anuales.
- Si se mantuviera esa tendencia, determinar la media anual del 2012 y aplicando los índices generales de variación estacional del mejor esquema (aditivo o multiplicativo), calcular la ocupación esperada para cada cuatrimestre del 2012.

2. La tabla siguiente muestra un estudio realizado en 5 años sobre el consumo eléctrico en una empresa, desglosado por trimestres:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2007	19	30	50	14
2008	20	30	50	15
2009	22	35	55	16
2010	25	37	65	17
2011	26	39	70	19

Calcular:

- Los índices normalizados de variación estacional según esquema multiplicativo.
- Calcular la tendencia mediante el método de mínimos cuadrados.
- Predecir el consumo eléctrico para cada trimestre del año 2012.

3. La tabla siguiente muestra un estudio realizado en 5 años sobre el porcentaje de agua en un embalse, desglosado por trimestres:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2007	60	30	58	25
2008	62	32	55	26
2009	65	37	60	27
2010	66	36	56	29
2011	67	38	58	30

Calcular:

- Los índices normalizados de variación estacional según esquema aditivo.
- Determinar la tendencia según el método de mínimos cuadrados.
- Predecir el porcentaje de agua para cada trimestre del año 2012.

4. La tabla siguiente muestra las ventas de cierto producto, en miles de unidades, durante 6 años, estudiada por cuatrimestres.

	Cuatrimestres		
Años	1º	2º	3º
2006	20	10	35
2007	20	15	40
2008	25	15	45
2009	30	20	50
2010	40	20	55
2011	40	25	55

- a) Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema multiplicativo.
- b) Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema aditivo.
- c) Determinar cuál de los dos esquemas es el más apropiado para este caso.
- d) Calcular una recta (por mínimos cuadrados) que explique la tendencia según las medias anuales.
- e) Si se mantuviera esa tendencia, determinar la media anual de ventas del 2012 y aplicando los índices generales de variación estacional del mejor esquema (aditivo o multiplicativo), calcular las ventas esperadas para cada cuatrimestre del 2012.

5. La tabla siguiente muestra un estudio realizado en 6 años sobre las medias de consumo de un producto por un panel de familias, desglosado por trimestres:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006	59	40	50	34
2007	60	35	50	25
2008	52	35	45	22
2009	47	27	45	19
2010	46	25	40	17
2011	45	20	35	15

Calcular:

- a) Los índices normalizados de variación estacional según esquema multiplicativo.
- b) Calcular la tendencia mediante el método de mínimos cuadrados.
- c) Predecir el consumo para cada trimestre del año 2012.

6. La tabla siguiente muestra un estudio realizado en 6 años sobre el número de personas que acuden a un multicine determinado, desglosado por trimestres:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006	60	70	68	35
2007	62	72	68	36
2008	65	77	69	37
2009	66	78	70	39
2010	66	78	70	40
2011	67	79	73	40

Calcular:

- Los índices normalizados de variación estacional según esquema aditivo.
- Determinar la tendencia según el método de mínimos cuadrados.
- Predecir el número de personas para cada trimestre del año 2012.

7. La tabla siguiente muestra un estudio realizado en 6 años sobre el número de reclamaciones sobre un servicio telefónico determinado, desglosado por trimestres:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006	60	70	68	35
2007	58	72	68	36
2008	58	74	67	36
2009	62	72	68	36
2010	58	75	71	39
2011	60	70	73	37

Calcular:

- Determinar si sería más conveniente utilizar el esquema aditivo o el multiplicativo para el cálculo de las variaciones estacionales.
- Calcular los índices de variación estacionales según el esquema escogido en el apartado anterior.
- Determinar la tendencia según el método de los mínimos cuadrados.
- Predecir el número de reclamaciones para cada trimestre del año 2012.

Tema 10

Los números índice

1. Introducción

Los números índice son indicadores que estudian las fluctuaciones de una cierta variable en función de un valor que se toma como base.

La finalidad de los números índice es caracterizar por un número único la variación de un conjunto de magnitudes o aspectos, considerados en dos momentos distintos del tiempo y/o el espacio.

Finalmente hemos de observar que es imprescindible, hablando de números índice, hacer referencia al término base, pues sin ello carece de sentido el indicador.

Los números índice pueden ser simples, cuando en ellos interviene una única variable, o complejos, cuando intervienen varias variables conjuntamente.

2. Índices simples

Si sólo se considera una magnitud variable, la relación entre sus valores en dos situaciones diferentes constituye un índice simple. Si nuestra variable es G , tomando dos valores en momentos diferentes, “ t ” y “ 0 ” (siendo “ 0 ” el valor que se toma como base), el índice correspondiente a G en el momento “ t ” respecto a su momento “ 0 ” se denota por I_0 y se calcula de la siguiente forma:

$$I_0 = \frac{G_t}{G_0} * 100$$

Se multiplican por 100 ya que los números índice suelen representarse en tanto por ciento.

3. Índices compuestos

La gran utilidad de los números índice aparece cuando se emplean para resumir en una sola serie las fluctuaciones de un conjunto de variables relacionadas entre sí.

Para que los números índice sean representativos ha de ponderarse cada uno de sus componentes en función de la importancia que tengan en el total.

Suele admitirse que si el índice que se trata de elaborar es de precios, se utiliza como ponderación las cantidades consumidas o producidas en el período, y si el índice es de cantidades, se utiliza como ponderación los precios. Dentro de esta norma genérica existen varios tipos de índices, entre los que destacan los de Laspeyres, Paasche y Fisher.

3.1. Índice de Laspeyres

Siguiendo la regla anterior, el índice de Laspeyres toma como ponderación las cantidades cuando se quiere calcular un índice de precios, y los precios cuando se quiere calcular un índice de cantidad. Si llamamos L_p al índice de precios Laspeyres y L_Q al índice de cantidad de Laspeyres, se pueden calcular como sigue:

$$L_p = \frac{\sum P_t \cdot Q_0}{\sum P_0 \cdot Q_0} = \frac{P_{1t} \cdot Q_{10} + P_{2t} \cdot Q_{20} + \dots + P_{nt} \cdot Q_{n0}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

$$L_Q = \frac{\sum P_0 \cdot Q_t}{\sum P_0 \cdot Q_0} = \frac{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

Tiene el inconveniente de admitir que los pesos establecidos para el período base se mantienen constantes en períodos sucesivos. Tiene, en cambio, la ventaja de que las ponderaciones para el año base son relativamente fáciles de obtener.

3.2. Índice de Paasche

Es análogo al anterior, pero utiliza como ponderaciones los datos correspondientes al período considerado. Igualmente, habrá dos índices de Paasche, que denotaremos por P_p (precios) y P_Q (cantidad).

$$P_p = \frac{\sum P_t \cdot Q_t}{\sum P_0 \cdot Q_t} = \frac{P_{1t} \cdot Q_{1t} + P_{2t} \cdot Q_{2t} + \dots + P_{nt} \cdot Q_{nt}}{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}$$

$$P_Q = \frac{\sum P_t \cdot Q_t}{\sum P_t \cdot Q_0} = \frac{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}{P_{1t} \cdot Q_{10} + P_{2t} \cdot Q_{20} + \dots + P_{nt} \cdot Q_{n0}}$$

3.3. Índice de Fisher

También llamado “fórmula ideal” es una media geométrica de los dos anteriores. Así,

$$F_p = \sqrt{L_p * P_p}$$

$$F_Q = \sqrt{L_Q * P_Q}$$

4. Utilidades de los números índice

Los números índice complejos pueden aplicarse a todo fenómeno en el que intervienen varias variables cuando interese conocer su evolución resumida en una sola serie. Los números índice se suelen aplicar en el cálculo de: Índices de Precios al Consumo (IPC), índices de producción, índices de salarios, índices de comercio exterior, índices de cotizaciones en bolsa, etc. El más conocido es el del IPC.

Ejercicios tema 10

1. El cuadro siguiente expresa los costes salariales medios por persona y hora trabajada en 10 años (en euros), junto con el total de horas trabajadas (en miles) en la empresa en esos años.

Período	Coste / hora	Horas
2000	7,00	1.200
2001	7,50	1.400
2002	7,70	1.100
2003	8,00	1.600
2004	8,10	1.500
2005	8,20	1.300
2006	8,50	1.600
2007	8,60	1.800
2008	8,90	1.700
2009	9,00	1.900

Calcular los índices de coste salarial (precios) y de horas trabajadas (cantidad) según Laspeyres, Paasche y Fisher, tomando como base el año 2005.

Con los mismos datos del ejercicio anterior, calcular los índices simples de coste salarial (precios) y de horas trabajadas (cantidad), tomando como base el año 2000.

2. Determinar los índices de precios y de cantidad de Laspeyres, Paasche y Fisher en 2004 para el conjunto de los siguientes productos tomando como base el año 2003:

Producto A			Producto B	
Año	Precio	Cantidad	Precio	Cantidad
2003	10,00	20.000	20,00	10.000
2004	12,00	21.000	25,00	8.000

3. Determinar los índices de precios y de cantidad de Laspeyres, Paasche y Fisher en 2009 para el conjunto de los siguientes productos tomando como base el año 2005:

Producto A			Producto B	
Año	Precio	Cantidad	Precio	Cantidad
2005	20,00	60.000	50,00	30.000
2009	29,00	71.000	65,00	48.000

4. Determinar los índices de precios y de cantidad de Laspeyres, Paasche y Fisher para el conjunto de los siguientes productos tomando como base el año 2007:

Producto A			Producto B	
Año	Precio	Cantidad	Precio	Cantidad
2007	10,00	30.000	5,00	20.000
2008	12,00	31.000	6,00	18.000
2009	14,00	35.000	6,00	22.000
2010	15,00	32.000	7,00	25.000
2011	15,00	34.000	8,00	23.000

Tema II

Contraste de hipótesis

1. Definición de inferencia estadística

La inferencia estadística pretende estimar resultados de la población a partir de los resultados obtenidos en una muestra representativa de ella.

El valor referido en la muestra se denomina **estadístico** y se representa con caracteres latinos (por ejemplo, la media es $\bar{X}$, la desviación típica es S , etc.). El valor que se supone en la población se denomina **parámetro** poblacional y se representa con caracteres del alfabeto griego (por ejemplo, la media es μ , la desviación típica es σ , etc.).

Esta inferencia estadística se realiza mediante los llamados test o contrastes de hipótesis, donde se pretende comprobar hasta qué punto un parámetro poblacional supuesto viene o no avalado por el estadístico obtenido en una muestra de esa población.

2. Las hipótesis

Una hipótesis estadística es una suposición relativa a una o varias poblaciones que puede ser o no cierta. Las hipótesis estadísticas se contrastan con los resultados obtenidos de las muestras para determinar si se aceptan o se rechazan como ciertas, en función de esos resultados obtenidos.

Ello hace que podamos hablar de dos tipos de errores:

- Error tipo I: rechazar la hipótesis cuando es cierta.
- Error tipo II: aceptar la hipótesis cuando es falsa.

	Hipótesis cierta	Hipótesis falsa
Se acepta la hipótesis		Error tipo II
Se rechaza la hipótesis	Error tipo I	

La probabilidad de error de tipo I y tipo II puede disminuirse aumentando el tamaño de la muestra.

Los pasos necesarios para realizar un contraste de hipótesis relativo a un parámetro son:

1. Establecer una hipótesis nula en términos de igualdad:

$$H_0: \theta = \theta_0$$

siendo θ la variable que estamos contrastando en nuestra hipótesis.

2. Establecer una hipótesis alternativa, que puede hacerse de tres formas diferentes:

$$H_1: \theta \neq \theta_0$$

$$H_1: \theta < \theta_0$$

$$H_1: \theta > \theta_0$$

3. Elegir un nivel de significación: recordaremos del tema 6 que el nivel de significación hace referencia a la fiabilidad del resultado. Se suele representar por α , indicando $1 - \alpha$ el grado de confianza.
4. Elegir un estadístico para el contraste: dependiendo del objeto de estudio y sus características, utilizaremos como estadísticos la *Chi*-cuadrado, la *Z* normal, la *t*-Student, la *F* de Snedecor, etc. Esto lo analizaremos en los siguientes epígrafes.

El contraste de hipótesis consiste en comprobar cuál de las dos hipótesis es verdadera o, lo que es lo mismo, si se acepta o se rechaza la hipótesis nula.

3. Análisis inferencial univariable

Dependiendo del tipo de variable a analizar, métrica o no métrica, se utilizan diferentes pruebas de contraste de hipótesis.

Cuando se trata de variables no métricas (escalas nominales u ordinales), en el análisis descriptivo resaltamos que el estudio consistiría en el cálculo de la moda o de las frecuencias. En el caso de inferencia estadística, la prueba a utilizar será la *Chi*-cuadrado.

En caso de variables métricas, en el análisis descriptivo se analiza fundamentalmente la media y la desviación típica. En el análisis inferencial se utilizará la prueba *Z* y la prueba *t* de Student.

3.1. Prueba Chi-Cuadrado (X^2) para variables no métricas

Se utiliza cuando el objetivo es comparar una distribución de frecuencias obtenida en una muestra con la distribución de frecuencias esperada en la población. Por ejemplo, obtenemos en una muestra unas frecuencias en diferentes tramos de edad y pretendemos comprobar si esos datos corroboran o no una supuesta distribución de frecuencias igual en tramos de edad de la población.

La fórmula a aplicar es la siguiente:

$$X^2 = \sum \frac{(\theta_i - E_i)^2}{E_i}$$

Donde,

θ_i = cada uno de los resultados observados en la muestra.

E_i = cada uno de los resultados esperados en la población.

Una vez obtenido el valor, comprobaremos en la tabla de la distribución de X^2 si este lo supera o no (podemos comprobarlo en el apéndice). Si el valor obtenido es inferior al de las tablas, podremos aceptar la hipótesis nula (que será concluir que la distribución de frecuencias es igual a la obtenida en la muestra). Si es superior al de las tablas, rechazaremos la hipótesis nula y por ello concluiremos que la distribución de frecuencias en la población no es igual al obtenido en la muestra.

La forma de buscar el valor en las tablas implica, en primer lugar, determinar el nivel de significación α . En segundo lugar, calcular los grados de libertad, que se expresa como $k - 1$ grados de libertad, donde k es el número de valores que toma la variable (en nuestro caso, sería los diferentes segmentos de edad establecidos).

Ejemplo: Analizada una muestra de 500 personas en cierta región geográfica se ha obtenido la siguiente distribución de frecuencias en función de la actividad de los individuos:

Segmento	Nº personas
Estudiantes	100
Trabajadores cuenta ajena	250
Trabajadores cuenta propia	50
Jubilados y pensionistas	100

Se establece como hipótesis nula que la población tiene la siguiente distribución para cada segmento (obtenido de aplicar la proporción poblacional al total de 500 personas encuestadas):

Segmento	Frecuencia
Estudiantes	110
Trabajadores cuenta ajena	240
Trabajadores cuenta propia	60
Jubilados y pensionistas	90

Para decidir si aceptamos o no esta afirmación en función de nuestros resultados aplicamos en primer lugar la fórmula siguiente:

$$X^2 = \sum \frac{(\theta_i - E_i)^2}{E_i}$$

Segmento	$\theta_i - E_i$	$(\theta_i - E_i)^2 / E_i$
Estudiantes	10	0,91
Trabajadores cuenta ajena	-10	0,42
Trabajadores cuenta propia	10	1,67
Jubilados y pensionistas	-10	1,11
	SUMA	4,11

Por tanto, $X^2 = 4,11$.

Acudimos ahora a las tablas de la X^2 para comprobar si este resultado es inferior o superior. Para ello, estableceremos un nivel de significación, que podría ser $\alpha = 0,05$. Además, determinamos que los k-1 grados de libertad sería 3 (son cuatro los segmentos analizados, menos 1, igual a 3). Comprobamos que, según tablas, con un nivel de significación de 0,05 y 3 grados de libertad, X^2 es 7,81. Como el resultado que hemos obtenido es menor, podemos concluir que nuestro estudio muestral corrobora la proporción supuesta para la población en cada uno de esos segmentos, con un 95% de confianza $(1 - \alpha)$.

3.2. Prueba Z para variables métricas

Se utiliza en variables métricas (cuantitativas) para comparar la media obtenida en una muestra con la media esperada en la población.

Para utilizarla es necesario que se cumplan una de las siguientes condiciones:

- La distribución de la población sigue una distribución normal y la varianza poblacional es conocida.
- La distribución de la población sigue una distribución normal, se desconoce la varianza, pero el tamaño de la muestra es grande (se considera así si es mayor de 30).

En caso de conocer la varianza poblacional, aplicaríamos la siguiente fórmula:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Donde,

$\bar{X}$, = media muestral

μ = media poblacional

σ = desviación típica poblacional

n = tamaño de la muestra

Si la varianza poblacional se desconoce y el tamaño de la muestra es superior a 30, emplearemos esta otra fórmula:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Donde,

S = desviación típica muestral

De esta forma, el resultado obtenido se corresponde con una distribución normal (0,1) que podremos buscar en las tablas. Si el resultado obtenido es mayor que el de las tablas, rechazaremos la hipótesis nula. En caso contrario, la aceptaremos.

Se define un nivel de significación α . Pero si la hipótesis alternativa implica dos colas, el nivel de significación será $\alpha/2$. O sea, suponiendo la hipótesis

nula $H_0: \theta = \theta_0$, siendo θ la variable que estamos contrastando en nuestra hipótesis, la probabilidad en tablas de la hipótesis alternativa puede ser:

$$H_1: P(\theta \neq \theta_0) = \alpha/2$$

$$H_1: P(\theta < \theta_0) = \alpha$$

$$H_1: P(\theta > \theta_0) = \alpha$$

Ejemplo 1: En un estudio de mercados se propone como hipótesis contrastar que las personas comienzan a utilizar cierto perfume a la edad de 32 años. En una muestra de 121 personas se ha obtenido que la media de edad en la que comienzan a utilizarlo es de 30 años, con una varianza de 144.

La hipótesis nula es $H_0: \mu = 32$.

La hipótesis alternativa es $H_1: \mu \neq 32$.

Por tanto, la hipótesis alternativa implica dos colas: mayor que 32 o menor que 32.

Definido el nivel de significación $\alpha=0,05$, aplicamos la fórmula:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$Z = \frac{30 - 32}{12 / \sqrt{121}} = 1,833 \text{ (tomamos los valores en términos absolutos)}$$

Buscamos en las tablas de la distribución normal (las encontramos en los anexos), y al ser de dos colas con $\alpha/2$, o sea, 0,025. Este valor se busca dentro de las soluciones de la tabla y encontramos ese valor en la fila de 1,9 y la columna 0,06. Estos valores se suman, siendo $Z = 1,96$. Al haber obtenido un resultado menor, aceptamos la hipótesis nula y concluimos que la edad media en la que las personas comienzan a utilizar este perfume es de 32 años.

Ejemplo 2: Otra forma alternativa de plantear este ejercicio sería considerar que las personas que utilizan el perfume tienen, de media, al menos 32 años. De esta forma, las hipótesis quedarían:

$$\text{Hipótesis nula } H_0: \mu \geq 32$$

$$\text{Hipótesis alternativa } H_1: \mu < 32$$

El estudio se corresponde al caso de una cola. Con los mismos datos del ejercicio anterior tenemos lo siguiente:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$Z = \frac{30 - 32}{12 / \sqrt{121}} = 1,833 \text{ (tomamos los valores en términos absolutos)}$$

Buscamos en las tablas de la distribución normal y al ser de una cola con nivel de significación α , o sea, 0,05, encontramos ese valor en la fila 1,6 y entre las columnas 0,04 y 0,05 (el punto intermedio entre estas dos últimas es 0,045). Se suman esos valores, siendo $Z = 1,645$. Al haber obtenido un resultado mayor, rechazamos la hipótesis nula y concluimos que la edad media en la que las personas utilizan este perfume no es ≥ 32 años.

3.3. Prueba t para variables métricas

Es similar a la prueba Z pero se utiliza cuando no se dan las condiciones para aplicar una distribución normal: cuando se desconoce la varianza poblacional y el tamaño de la muestra es pequeño.

La fórmula a utilizar es la siguiente:

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

La utilización de las tablas de la distribución t de Student es similar a la ya comentada Z normal. Puede ser con una o dos colas, con lo que el grado de significación será α o $\alpha/2$, respectivamente.

Ejemplo 1: Supongamos el mismo ejemplo anterior, donde en un estudio de mercados se propone como hipótesis contrastar que las personas comienzan a utilizar cierto perfume a la edad de 32 años. En una muestra de 25 personas se ha obtenido que la media de edad en la que comienzan a utilizarlo es de 30 años, con una varianza de 100.

La hipótesis nula es $H_0: \mu = 32$

La hipótesis alternativa es $H_1: \mu \neq 32$

Por tanto, la hipótesis alternativa implica dos colas: mayor que 32 o menor que 32.

Definido el nivel de significación $\alpha=0,05$, aplicamos la fórmula:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$t = \frac{30 - 32}{12 / \sqrt{25}} = 1,00 \text{ (tomamos los valores en términos absolutos)}$$

Buscamos ahora en las tablas t de Student, con nivel de significación $\alpha/2$ ($0,05/2 = 0,025$) y k-1 grados de libertad ($25-1 = 24$), obteniendo como valor de $t = 2,064$. Al ser este mayor que el obtenido en la fórmula, podemos aceptar la hipótesis nula, con lo que la edad en la que las personas comienzan a utilizar este perfume es de 32 años.

Ejemplo 2: Siguiendo con el mismo ejemplo anterior, donde en un estudio de mercados se propone como hipótesis contrastar que las personas utilizan cierto perfume a partir de la edad de 32 años, en una muestra de 25 personas se ha obtenido que la media de edad en la que comienzan a utilizarlo es de 30 años, con una varianza de 100.

La hipótesis nula es $H_0: \mu \geq 32$.

La hipótesis alternativa es $H_1: \mu < 32$.

Por tanto, la hipótesis alternativa implica una cola.

Definido el nivel de significación $\alpha=0,05$, aplicamos la fórmula:

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$t = \frac{30 - 32}{12 / \sqrt{25}} = 1,00 \text{ (tomamos los valores en términos absolutos)}$$

Buscamos ahora en las tablas t de Student, con nivel de significación α (0,05) y k-1 grados de libertad ($25 - 1 = 24$), obteniendo como valor de $t = 1,711$. Al ser este mayor que el obtenido en la fórmula, podemos aceptar la hipóte-

sis nula, con lo que la edad en la que las personas utilizan este perfume es a partir de los 32 años.

4. Análisis inferencial bivariable

El análisis inferencial bivariable es el más utilizado en la práctica de la investigación comercial. Permite, por ejemplo, comprobar la existencia de relación entre dos variables analizadas (causa-efecto) o comprobar la coherencia de resultados entre dos muestras de la misma población.

En estos estudios podemos contrastar variables no métricas o variables métricas. En el caso de variables no métricas, el análisis se realizará mediante tablas cruzadas con el test X^2 . En caso de variables métricas, el análisis descriptivo lo realizaríamos, tal y como vimos en el tema 8, mediante el estudio de la covarianza, mientras que el análisis inferencial se realiza utilizando el llamado *test de medias*.

4.1. Test Chi-Cuadrado: tablas cruzadas

Cuando tratamos variables no métricas, puede resultar conveniente contrastar la hipótesis de que una de las variables depende de la otra. Por ejemplo, podemos analizar si el uso de cuatro marcas de perfume depende de tres diferentes segmentos de nivel de renta de las personas.

En estos casos se utiliza el test X^2 mediante tablas cruzadas. La fórmula a utilizar es la misma que la ya vista para inferencias univariadas:

$$X^2 = \sum \sum \frac{(\theta_{ij} - E_{ij})^2}{E_{ij}}$$

Donde,

θ_{ij} = resultado observado en la muestra, columna i, fila j.

E_{ij} = resultado esperado en la población, columna i, fila j.

En estos estudios la hipótesis nula será siempre considerar que no existe relación entre las variables.

Ejemplo: Se analiza una muestra de 500 personas determinando 3 niveles de renta (alta, media y baja) respecto al consumo de 4 marcas de perfume (A, B, C y D). Los resultados obtenidos son los siguientes:

Renta	Marca A	Marca B	Marca C	Marca D	Totales
Alta	70	50	40	20	180
Media	40	60	30	40	170
Baja	10	30	60	50	150
Total	120	140	130	110	500

Si no existiera relación entre la renta y el consumo de un determinado perfume, podríamos establecer una tabla con las frecuencias esperadas de la población extrapoliándola a un total de 500 individuos, realizando las siguientes operaciones:

En renta alta:

$$\frac{180}{500} = \frac{E_{AA}}{120} = \frac{E_{AB}}{140} = \frac{E_{AC}}{130} = \frac{E_{AD}}{110}$$

Donde,

E_{AA} = frecuencia esperada de renta alta que consumen la marca A.

E_{AB} = frecuencia esperada de renta alta que consumen la marca B.

E_{AC} = frecuencia esperada de renta alta que consumen la marca C.

E_{AD} = frecuencia esperada de renta alta que consumen la marca D.

En renta media:

$$\frac{170}{500} = \frac{E_{MA}}{120} = \frac{E_{MB}}{140} = \frac{E_{MC}}{130} = \frac{E_{MD}}{110}$$

Donde,

E_{MA} = frecuencia esperada de renta media que consumen la marca A.

E_{MB} = frecuencia esperada de renta media que consumen la marca B.

E_{MC} = frecuencia esperada de renta media que consumen la marca C.

E_{MD} = frecuencia esperada de renta media que consumen la marca D.

En renta baja:

$$\frac{150}{500} = \frac{E_{BA}}{120} = \frac{E_{BB}}{140} = \frac{E_{BC}}{130} = \frac{E_{BD}}{110}$$

Donde,

E_{BA} = frecuencia esperada de renta baja que consumen la marca A.

E_{BB} = frecuencia esperada de renta baja que consumen la marca B.

E_{BC} = frecuencia esperada de renta baja que consumen la marca C.

E_{BD} = frecuencia esperada de renta baja que consumen la marca D.

Llevando los resultados a una tabla, tenemos que las frecuencias esperadas en la población son las siguientes:

Renta	Marca A	Marca B	Marca C	Marca D	Totales
Alta	43,2	50,4	46,8	39,6	180
Media	40,8	47,6	44,2	37,4	170
Baja	36	42	39	33	150
Total	120	140	130	110	500

Ahora aplicamos la fórmula:

$$X_2 = \sum \sum \frac{(\theta_{ij} - E_{ij})^2}{E_{ij}}$$

El resultado de aplicar la fórmula a cada celda se expresa en la tabla siguiente:

Renta	Marca A	Marca B	Marca C	Marca D	Totales
Alta	16,6259259	0,00317	0,98803	9,70101	
Media	0,01568627	3,23025	4,56199	0,18075	
Baja	18,7777778	3,42857	11,3077	8,75758	
Total					77,5784401

Ahora buscamos en la tabla de la distribución X^2 , con $(n - 1)(m - 1)$ grados de libertad (o sea, $2 \times 3 = 6$), suponiendo un nivel de significación de 0,05. La hipótesis nula H_0 : es que no existe relación entre las variables. El valor en las tablas, con $\alpha = 0,05$ y 6 grados de libertad es 12,59. Este valor es claramente inferior al obtenido con la fórmula (77,58). Por ello, rechazamos la hipótesis nula y podemos concluir que sí existe relación entre la renta y el consumo de una marca específica de perfume.

Si ahora calculamos los porcentajes dentro de cada nivel de renta en cuanto al consumo de cada una de las marcas, tendríamos lo siguiente:

Renta	Marca A	Marca B	Marca C	Marca D	Totales
Alta	38,89%	27,78%	22,22%	11,11%	100
Media	23,53%	35,29%	17,65%	23,53%	100
Baja	6,67%	20,00%	40,00%	33,33%	100

Con ello, apreciamos que los individuos de renta alta prefieren la marca A, los de renta media prefieren la marca B y los de renta baja prefieren la marca C.

Ahora bien, en este tipo de investigaciones hemos de ser cautelosos. Hemos analizado la influencia de una variable en otra, pero ocurre en ocasiones que existen otras causas que explican esta influencia. En nuestro ejemplo, podría deberse al tipo de establecimiento donde cada clase social realiza sus compras, o bien al tramo de edad más frecuente en cada clase social. Podría ocurrir que las personas de renta alta realizan sus compras en comercios especializados donde no se vende la marca C y D, por ejemplo. Esto supondría un sesgo en la información, ya que pretendemos establecer la relación entre dos variables, cuando en realidad pueden existir otras que también influyan y enmascaren la información.

4.2. Test de medias

Se pueden plantear dos cuestiones de análisis:

- Comprobar si las diferencias en las medias obtenidas en dos muestras diferentes de una población se deben a errores muestrales.
- Comprobar si existen diferencias en las medias obtenidas en una misma muestra realizadas en dos momentos diferentes del tiempo.

El primer caso se resuelve mediante el análisis de *test de medias para muestras independientes*. El segundo caso, mediante el *test de medias para muestras relacionadas*.

4.2.1. Test de medias para muestras independientes

Se pretende comprobar si la diferencia de medias encontrada en dos muestras independientes de la misma población corresponde o no a los errores

muestrales o bien se puede establecer que la diferencia en las medias muestrales supone que existan diferencias en las medias poblacionales. Un ejemplo sería comprobar si existen diferencias de consumo entre hombres y mujeres respecto a un producto concreto. Elegida una muestra, se encuentra una media de consumo de hombres determinada. Elegida otra muestra, se calcula el consumo medio del producto entre las mujeres. Las diferencias entre una y otra media puede indicar diferencias en las medias de consumo de los hombres y mujeres de la población, o bien se deben a errores muestrales aceptables.

La hipótesis nula será

$$H_0: \mu_1 = \mu_2$$

En estos casos, se aplica el estadístico Z normal, si es conocida la varianza poblacional o bien el tamaño de muestra es grande (al menos 30 individuos), y se utiliza t de Student si no cumple ninguno de estos requisitos.

La fórmula a utilizar en caso de distribuciones Z es la siguiente:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Una vez calculada, compararemos este valor con el de la tabla de Z normal. Si este último es mayor, aceptamos la hipótesis nula y por tanto no existen diferencias significativas de medias. En caso contrario, rechazamos la hipótesis nula.

Si utilizamos la t de Student, la fórmula es la misma, pero buscaríamos en la tabla correspondiente, con $n_1 + n_2 - 2$ grados de libertad.

Ejemplo 1: En una muestra de 81 hombres encontramos que la media de consumo de cierta marca de leche es de 16, con una varianza de 25. En otra muestra de 85 mujeres, la media de consumo de esa marca de leche es de 17, con una varianza de 24 ¿Podemos concluir con ello que existen diferencias de consumo entre hombres y mujeres respecto a esa marca de leche?

Solución: La hipótesis nula será que la media de consumo es igual entre hombres y mujeres.

Aplicamos la fórmula:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$Z = \frac{16 - 17}{\sqrt{\frac{25}{81} + \frac{24}{85}}}$$

$$Z = \frac{-1}{\sqrt{0,3086 + 0,2824}} = 1,30 \text{ (tomamos valores absolutos)}$$

Buscando en la tabla Z normal, para un nivel de significación $\alpha = 0,05$ (al ser de dos colas, puesto que $H_1: \mu_1 \neq \mu_2$, buscaremos con nivel de significación $\alpha/2 = 0,025$), este valor corresponde a $Z = 1,96$. Como el valor obtenido es menor, aceptamos la hipótesis nula y por tanto concluimos que no existen diferencias en las medias de consumo de esa marca de leche entre hombres y mujeres.

Ejemplo 2: Elegida una muestra de 20 personas solteras se comprueba que la media de consumo de cerveza es de 17 botellines al mes, con una varianza de 30. Otra muestra de 22 personas casadas ofrece una media de consumo de 12 botellines al mes, con una varianza de 40 ¿Podemos concluir que el consumo medio de botellines de cerveza es distinto entre las personas solteras que entre las casadas?

Solución: Estamos ante un caso de t de Student al ser la muestra pequeña y desconocer la varianza poblacional. La hipótesis nula será que el consumo de cerveza es igual entre las personas solteras y casadas, frente a la hipótesis alternativa que sería considerar que el consumo de cerveza es distinto en las personas solteras que en las casadas.

Aplicamos la fórmula siguiente:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{17 - 12}{\sqrt{\frac{30}{20} + \frac{40}{22}}}$$

$$t = \frac{5}{\sqrt{1,5 + 1,82}} = 2,74$$

Buscamos ahora en la tabla t de Student, con 40 grados de libertad ($n_1 + n_2 - 2$) y $\alpha = 0,05$. El valor de t es 1,684. Como éste es menor que el calculado, rechazamos la hipótesis nula y consideramos que existen diferencias de consumo de botellines de cerveza entre las personas solteras y las casadas.

4.2.2. Test de medias para muestras relacionadas

Es el caso de tomar de una misma muestra dos medias en diferentes momentos del tiempo. Sería habitual cuando queremos comprobar, por ejemplo, si la media de consumo de cierto producto ha experimentado una variación significativa después de realizada cierta campaña publicitaria o cualquier otra acción comercial.

La hipótesis nula se plantea como la no existencia de diferencias entre los dos parámetros de la población.

La fórmula a aplicar se basa en el empleo de la t de Student con $n - 1$ grados de libertad:

$$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

Donde,

$\bar{d}$ = media de las diferencias encontradas en las medidas tomadas a la muestra en diferentes momentos del tiempo.

S_d = desviación típica de las diferencias.

n = tamaño de la muestra.

Ejemplo: En una muestra de 121 individuos se calcula la media de consumo de cierto producto antes y después de bajar el precio y se ha obtenido que la media de las diferencias en las respuestas de los encuestados es de 4,2. La

desviación típica de las diferencias es de 2,3 ¿Podemos aceptar que existen diferencias de consumo antes y después de bajar los precios?

Solución: Aplicamos la fórmula:

$$t = \frac{4,20}{2,3\sqrt{121}} = 20,09$$

Buscamos ahora en las tablas de t de Student, con 120 grados de libertad (k-1) y un nivel de significación de 0,05. El valor en tablas de t es de 1,658. Como el obtenido con la fórmula es muy superior podemos decir que se rechaza la hipótesis nula, esto es, sí hay diferencias significativas en el consumo al bajar los precios.

Ejercicios tema 11

1. Analizada una muestra de 1.000 personas se ha obtenido la siguiente distribución de frecuencias en función de la edad de los individuos:

Segmento	Nº personas
Niños	200
Jóvenes	250
Mediana edad	400
Ancianos	150

Según datos sociográficos la población tiene la siguiente distribución para cada segmento:

Segmento	Frecuencia
Niños	20%
Jóvenes	24%
Mediana edad	42%
Ancianos	14%

¿Podemos afirmar que, según nuestro estudio, se confirma esa distribución poblacional, con un nivel de confianza del 99%?

2. En un estudio de mercado se propone como hipótesis contrastar que las personas comienzan a utilizar cierto tipo de calzado a la edad de 21 años. En una muestra de 81 personas se ha obtenido que la media de edad en la que comienzan a utilizarlo es de 22 años, con una varianza de 121. ¿Podemos determinar que la edad para comenzar a utilizarlo es de 21 años, con un nivel de confianza del 95%?

3. En un estudio de mercado pretendemos determinar si la edad media en que los jóvenes comienzan a utilizar el teléfono móvil es a partir de los 12 años. Hemos realizado una encuesta a 400 jóvenes y el resultado obtenido es que comienzan a utilizarlo a los 11 años, con una varianza de 36. ¿Nuestro estudio corrobora la afirmación de que la edad en que los jóvenes comienzan a utilizar el teléfono móvil es a los 12 años? Realizarlo con un nivel de confianza del 95%.

- 4.** En un estudio de mercados se propone como hipótesis contrastar que las personas comienzan a utilizar una determinada marca de champú a la edad de 36 años. En una muestra de 20 personas se ha obtenido que la media de edad en la que comienzan a utilizarlo es de 34 años, con una varianza de 121. ¿Podemos corroborar la suposición de que la edad en que comienzan a utilizarlo es a los 36 años, con un nivel de confianza del 95%?
- 5.** En un estudio de mercados se propone como hipótesis contrastar que los clientes de cierto comercio de moda tienen una edad de 35 años o más. En una muestra de 25 personas se ha obtenido que la media de edad en la que acuden a este tipo de establecimiento es a partir de 37 años, con una varianza de 81. Según nuestro estudio ¿podemos concluir que, efectivamente, la edad de la clientela de este establecimiento es a partir de 35 años, con un nivel de confianza del 95%?
- 6.** Se analiza una muestra de 1.000 personas determinando 3 segmentos de edad (jóvenes, adultos y ancianos) respecto al consumo de 3 marcas de galletas (A, B y C). Los resultados obtenidos son los siguientes:

Edad	Marca A	Marca B	Marca C	Totales
Jóvenes	90	50	240	380
Maduros	140	50	80	270
Ancianos	80	200	70	350
Total	310	300	390	1.000

¿Podemos afirmar que la edad influye en la elección de la marca de galletas, con un nivel de confianza del 95%?

- 7.** En una muestra de 121 hombres encontramos que la media de consumo de cierta marca de gel de ducha es de 25 unidades al año, con una varianza de 36. En otra muestra de 115 mujeres, la media de consumo de esa marca de gel de ducha es de 27 unidades al año, con una varianza de 32. ¿Podemos concluir con ello que existen diferencias de consumo entre hombres y mujeres respecto a esa marca de gel de ducha, con un nivel de confianza del 95%?
- 8.** Elegida una muestra de 15 personas menores de 30 años se comprueba que la media de asistencia al cine es de 7 veces al mes, con una varianza de 5. Otra muestra de 17 personas mayores de 30 años ofrece una media de asistencia al cine de 2 veces al mes, con una varianza de 3. ¿Podemos con-

cluir que la asistencia media al cine es distinta entre los menores de 30 años que entre los mayores de 30 años?

9. En una muestra de 100 individuos se calcula la media de consumo mensual de cierto producto antes y después de una campaña publicitaria y se ha obtenido que la media de las diferencias en las respuestas de los encuestados es de 2,2. La desviación típica de las diferencias es de 0,3. ¿Podemos aceptar que existen diferencias de consumo antes y después de la campaña publicitaria?

Tema 12

El informe de investigación comercial

1. Introducción

El Informe final es el último paso en el proceso de investigación. Es un documento escrito que tiene el propósito de dar a conocer algo: presentando hechos y datos obtenidos y elaborados, su análisis e interpretación, indicando los procedimientos utilizados y llegando a ciertas conclusiones y recomendaciones. Su objetivo es el de comunicar los resultados de una investigación.

En la redacción del informe final de una investigación de mercados se debe tener presente el fin para el que será utilizado, que podemos enumerar a continuación:

- El informe sirve para tomar decisiones.
- El valor del informe radica en su utilidad.
- La redacción y presentación del informe es extremadamente importante.
- El informe tiene dos formas, una de presentación oral ante el cliente, y otra, la de un informe escrito que se presenta, revisa y consulta.
- Su valoración por el cliente va a depender de cómo se ha analizado la información y cómo se presenta.
- El informe debe ser útil, aprovechable y verosímil y debe estar dirigido a los que lo utilizan o aplican.

2. Tipos de informe

En general se suelen distinguir cuatro tipos de informes, considerando como criterio de clasificación los destinatarios y fines de la investigación:

- **Informes científicos:** van destinados a hombres de ciencia, consecuentemente competentes en el tema que trata la investigación; en

este caso, el lenguaje es riguroso y no hay limitaciones en el uso de tecnicismos; estos informes pertenecen a la categoría de “memorias científicas”.

- **Informes técnicos:** destinados a las organizaciones públicas o privadas que han encargado el estudio o investigación; en este caso, manteniendo el máximo rigor, se procurará que el informe sea accesible a los destinatarios, que no siempre dominan toda la jerga propia de la sociología, antropología, psicología social, etc.
- **Informes de divulgación:** se trata de estudios destinados al público en general; por consiguiente, deben ser escritos en un lenguaje accesible a una persona de cultura media.
- **Informes mixtos:** suelen estar destinados a una organización, al mismo tiempo que se dan a conocer al público en general.

3. Formato del informe

No existe un formato específico que sea adecuado para todas las situaciones. Un trabajo de investigación no está concluido hasta tanto haya sido escrito el informe. La hipótesis más brillante, el estudio más cuidadosamente preparado y realizado, los resultados más sorprendentes son de escaso valor a menos que sean comunicados a otros.

Se presentan dos informes: uno escrito, prolijo, con toda la documentación y tan extenso como sea necesario. El otro informe se realiza en forma de imágenes, cuadros, etc., para ser presentado oralmente.

La siguiente guía es aceptada generalmente como el formato básico para la mayor parte de los proyectos de investigación:

1. Portada
2. Tabla de contenido
3. Índice de las tablas (o figuras, gráficas, etc.)
4. Resumen gerencial
 - a. Objetivos
 - b. Resultados
 - c. Conclusiones
 - d. Recomendaciones

- 5. Cuerpo
 - a. Introducción
 - b. Resultados
 - c. Limitaciones
- 6. Conclusiones y recomendaciones
- 7. Apéndice
 - a. Plan muestral
 - b. Formatos de recolección de datos
 - c. Tablas de apoyo no incluidas en el cuerpo

4. Estructura de los informes

En lo que concierne a la estructura de los informes, esta tiene una secuencia lógica que, en términos generales, explica de qué se trata, qué se hizo, cómo se hizo y cuáles son las conclusiones. Cualquiera que sea la longitud o la índole de los informes, estos tienen ciertos elementos comunes que constituyen su estructura básica. Una forma de hacerlo más o menos universalmente admitida es la siguiente:

1. **Portada:** La portada debe contener un título que resuma la esencia del estudio, fecha, nombre de la organización que está presentando el informe y nombre de la organización a quien va dirigido e informe. Si el informe es confidencial, los individuos que van a recibirlo deben estar incluidos en esta página.
2. **Tablas de contenido:** La tabla de contenido enumera en forma secuencial los temas cubiertos en el informe, junto con sus referencias de páginas. Su propósito es ayudar a los lectores a encontrar secciones específicas del informe que son de mayor interés para ellos.
3. **Índice de tablas:** Este índice enumera los títulos y números de páginas de todas las ayudas visuales. Esta tabla se puede ubicar en la misma página de la tabla de contenido o en una página separada.
4. **Resumen gerencial:** El resumen gerencial es una presentación concisa y exacta de los aspectos fundamentales del informe. Esta sinopsis debe tener una extensión de una o dos páginas. Puesto que muchos

ejecutivos leen únicamente el resumen gerencial, es importante que esta sección sea concisa y que esté escrita en forma adecuada. El resumen proporciona, a quien toma decisiones, los resultados de la investigación que presentan un mayor impacto en la decisión que se tiene que tomar. Este resumen debe incluir:

- Objetivos del proyecto de investigación.
- Naturaleza del problema de decisión.
- Resultados clave.
- Conclusiones (opiniones e interpretaciones basadas en la investigación).
- Recomendaciones para la acción.

5. Cuerpo del Informe: Los detalles del proyecto de investigación se encuentran en el cuerpo del informe. Esta sección incluye: Introducción, metodología, resultados y limitaciones.

6. Introducción: El propósito de la introducción es proporcionar al lector la información básica necesaria para comprender el resto del informe. La naturaleza de la introducción está condicionada por la diversidad de la audiencia y su familiarización con el proyecto de investigación. Cuando más diversa sea la audiencia, más extensa será la introducción. La introducción debe explicar claramente la naturaleza del problema de decisión y el objeto de la investigación. La información básica debe relacionarse con el producto o servicio involucrado, y las circunstancias que rodean el problema de decisión. Debe revisarse la naturaleza de cualquier tipo de investigación anterior al problema.

7. Metodología: El propósito de la sección de metodología es describir la naturaleza del diseño de investigación, plan muestral y procedimientos de recolección y análisis de datos. Esta es una sección muy difícil de escribir. Se debe dar el suficiente detalle para que el lector pueda apreciar la naturaleza de la metodología empleada, pero la presentación no debe ser excesiva o monótona. Se debe evitar el uso de la jerga técnica. La sección de metodología debe decir al lector si el diseño era exploratorio o concluyente. Deben explicarse las fuentes de datos, secundarias o primarias. Debe especificarse la naturaleza del método de recolección de datos, comunicación y observación. El lector necesita saber a quién se incluyó en la muestra, tamaño de la

muestra y naturaleza del procedimiento muestral. Esta sección está diseñada para resumir los aspectos técnicos del proyecto de investigación en un estilo que sea comprensible por una persona que no es técnica; debe desarrollar confianza en la calidad de los procedimientos empleados. Se deben minimizar los detalles técnicos y colocarlos en un apéndice para quienes deseen un análisis metodológico más detallado.

- 8. Resultados:** Está compuesto por los resultados de la investigación, los cuales deben organizarse alrededor de los objetivos de la misma y de las necesidades de información. Esta presentación debe comprender una exposición lógica de la información, como si se fuera a contar una historia. El informe de los hallazgos debe tener un punto de vista definitivo y una secuencia lógica; no es simplemente la presentación de una serie interminable de tablas. Más bien, se requiere la organización de los datos en un flujo lógico de información a propósito de la toma de decisiones.
- 9. Limitaciones:** Cada proyecto de investigación tiene limitaciones que es necesario comunicar de una forma clara y concisa. En este proceso, el investigador debe evitar comentar las debilidades menores del estudio. El propósito de esta sección no es disminuir la calidad del proyecto de investigación, sino permitir que el lector haga un juicio sobre la validez de los resultados del estudio. Las limitaciones en un proyecto de investigación de mercados generalmente involucran las inexactitudes del muestreo y la no respuesta, así como posibles debilidades metodológicas. La redacción de la sección de conclusiones y recomendaciones está afectada por las limitaciones reconocidas y aceptadas del estudio. Es responsabilidad profesional del investigador informar claramente al lector sobre estas limitaciones.
- 10. Conclusiones y recomendaciones:** Las conclusiones y las recomendaciones deben fluir en una forma lógica a partir de la presentación de los resultados. Las conclusiones deben relacionar en forma clara los hallazgos de la investigación con las necesidades de información y, en base a ello, hacer las recomendaciones para la acción.
- 11. Apéndice:** El apéndice proporciona un espacio para el material que no es esencial en el cuerpo del informe. Este material es más especializado y más complejo que el material presentado en el informe principal y se incluye para satisfacer las necesidades del lector. Con

frecuencia, contendrá copias de los formatos de recolección de datos, detalles del plan muestral, estimaciones del error estadístico, instrucciones para el entrevistador y tablas estadísticas detalladas asociadas con el proceso de análisis de datos.

5. Presentación de los datos

Cuando llega el momento de presentar los informes de investigación, las dos ayudas que más se utilizan son las tablas y las figuras. Estas hacen el informe menos complejo y más fácil de leer y comprender. Las ayudas gráficas mejoran la apariencia física.

Es conveniente seguir las siguientes recomendaciones:

- Es mejor usar una ilustración dentro del texto si el lector necesita referirse a éste mientras lee el informe.
- Si la información es suplementaria o demasiada larga puede colocarse en un apéndice.
- Siempre debe hacerse una introducción de la ilustración antes de presentarla al lector.
- No analice detalles minuciosos de la ilustración; los lectores lo encontrarán complejo y redundante.

Todas las ayudas gráficas deben incluir los siguientes elementos:

- Número de la tabla o figura.
- Títulos: indica el contenido de la tabla o figura.
- Rotulaciones: el encabezamiento contiene las leyendas o enunciados de las columnas en una tabla, mientras que los márgenes contienen las rotulaciones de las filas.
- Notas de pie de página.

Los datos pueden presentarse en forma de tablas o gráficas. Las gráficas pueden mejorar una presentación centrando la atención en los puntos importantes que no pueden explicarse claramente en las tablas. Son medios rápidos y atractivos de presentar números, tendencias y relaciones.

Soluciones de los ejercicios

Soluciones ejercicios. Tema 6

Ejercicio I

a) Se trata de una población finita:

$$n = \frac{4Np(l-p)}{K^2(N-l) + 4p(l-p)}$$

$$n = \frac{4 * 25.000 * 0,5 * (1 - 0,5)}{0,02^2 * (25.000 - 1) + 4 * 0,5 * (1 - 0,5)}$$

$$n = 25.000 / 10,9996 = 2.273,554$$

El tamaño de la muestra debe ser de 2.274 encuestados.

b)

$$S = \sqrt{\frac{(N-n) p(l-p)}{(N-l) n}}$$

$$S = \sqrt{\frac{(25.000 - 3.000) 0,5 * (1 - 0,5)}{(25.000 - 1) 3.000}}$$

$$S = \sqrt{\frac{(22.000) 0,25}{(24.999) 3.000}}$$

$$S = \sqrt{0,8800352 * 0,0000833}$$

$$S = 0,008561946$$

Con un grado de confianza del 95%, el error muestral es 2S, por tanto del 1,71%.

c) Hemos de tener en cuenta el error muestral, que en este caso es del 1,71%.

Por ello, al encontrar en la muestra un 12% de menores de 15 años, para extrapolar esta información a la población debemos aplicar un margen de $\pm 1,71\%$. Como el 1,71% del 12% es 0,21%, podemos decir que el porcentaje de menores de 15 años en la población oscilará entre 11,79% y 12,21%.

En este sentido, pasando esos porcentajes a cifras, sabiendo que la población es de 25.000 personas, podemos decir que el número de menores de 15 años oscila entre 2.948 (11,79% de 25.000) y 3.052 (12,21% de 25.000). O lo que es lo mismo, son 3.000 ± 52 personas.

d) Recordaremos que el método sistemático parte del censo ordenado y se elige a los individuos tomando una k-ésima unidad de la población, siendo:

$$K = N / n$$

En nuestro caso:

$$K = 25000 / 3000 = 8,3333$$

Por ello, elegido el primer individuo, el siguiente será el que ocupe 8 puestos más adelante en esa ordenación (que podría ser por orden alfabético, por ejemplo). Contaremos 8 más a partir de este último y se elige el siguiente. De esta forma elegiremos los 3.000 individuos de la población.

Es fundamental contar con el censo de la población, obtenido de alguna fuente externa, que aparecerá ordenado por algún criterio, que en estos casos suele ser por orden alfabético del primer apellido. Podríamos cambiar ese orden, pero resultaría un exceso de trabajo que posiblemente no reporte ninguna mejora en la selección. En cambio, si el objetivo es encuestar sólo a una parte de esa población con características psicosociales determinadas, sí sería conveniente reordenar el censo.

Ejercicio 2

a) Se trata de una población no finita, al superar 100.000 elementos.

El tamaño de la muestra debe ser de 1.111 encuestados.

b) Con un grado de confianza del 95%, el error muestral es 2S, por tanto del 1,41%.

c) Hemos de tener en cuenta el error muestral, que en este caso es del 1,41%.

Por ello, al encontrar en la muestra un 25% de fumadores, para extrapolar esta información a la población debemos aplicar un margen de $\pm 1,41\%$. Como el 1,41% del 25% es 0,36%, podemos decir que el porcentaje de fumadores en la población oscilará entre 24,64% y 25,36%.

En este sentido, pasando esos porcentajes a cifras, sabiendo que la población es de 250.000 personas, podemos decir que el número de fumadores oscila entre 61.600 (24,64% de 250.000) y 63.400 (25,36% de 250.000). O lo que es lo mismo, son 60.000 ± 400 personas.

d) El muestreo por conglomerados consiste en elegir al azar grupos para tomar luego una submuestra dentro de esos grupos. En nuestro caso, dada una población determinada, podríamos partir de una relación de las calles de la ciudad.

A cada calle se le asigna un número de orden. A partir de ahí, extraemos al azar una serie de números que representarán las calles donde seleccionaremos la muestra. Podríamos incluso sortear los números de vivienda en cada calle (estaríamos en un muestreo de dos etapas).

Una vez elegidas las calles y los números de vivienda, se entrevistarán a los individuos hasta llegar al número decidido, que en este caso son 5.000 personas.

Este método por conglomerados puede realizarse combinado con otros métodos, como por ejemplo el sistemático, por estratos o aleatorio simple. Incluso podríamos combinarlo con métodos no aleatorios.

Soluciones ejercicios. Tema 7

Ejercicio I

En primer lugar, hemos de calcular el rango, que es la diferencia entre el mayor valor y el menor:

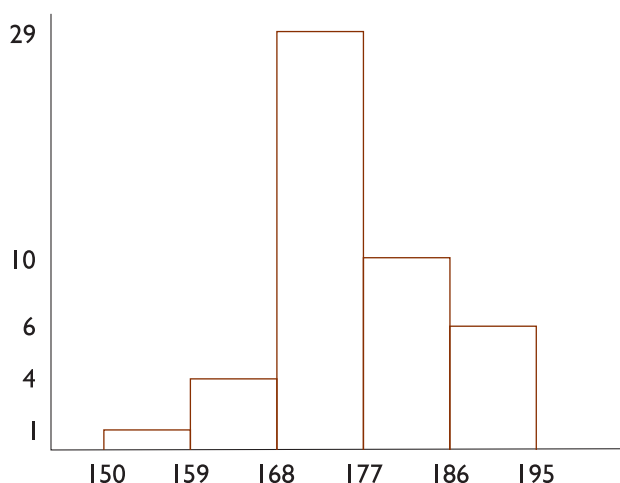
$$\text{Rango} = 195 - 150 = 45$$

Como queremos determinar 5 clases de la misma amplitud, dividiremos el rango entre el número de clases para calcular esa amplitud:

$$\text{Amplitud} = \text{Rango} / \text{N}^\circ \text{ clases} = 45 / 5 = 9$$

Estatura	Frecuencia
[150 – 159]	1
[159 – 168]	4
[168 – 177]	29
[177 – 186]	10
[186 – 195]	6

Ahora, representamos esos datos en un histograma:



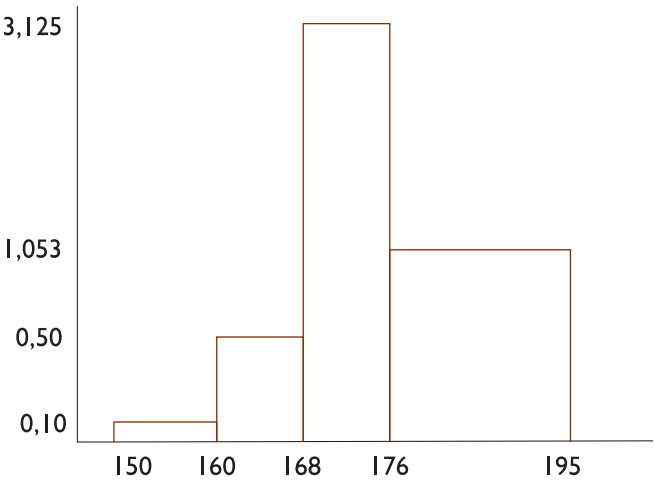
Ejercicio 2

Estatura	Frecuencia
[150 – 160]	1
[160 – 168]	4
[168 – 176]	25
[176 – 195]	20

Para calcular el histograma, dado que son clases de diferente amplitud, hemos de calcular las alturas (h_i), dividiendo la frecuencia de cada intervalo entre su amplitud.

Estatura	Frecuencia	Altura
[150 – 160]	1	0,10
[160 – 168]	4	0,50
[168 – 176]	25	3,125
[176 – 195]	20	1,053

En el histograma representaremos las clases y su altura:



Ejercicio 3

En primer lugar, hemos de calcular el rango, que es la diferencia entre el mayor valor y el menor:

$$\text{Rango} = 35.000 - 10.000 = 25.000$$

Como queremos determinar 5 clases de la misma amplitud, dividiremos el rango entre el número de clases para calcular esa amplitud:

$$\text{Amplitud} = \text{Rango} / \text{N}^\circ \text{ clases} = 25.000 / 5 = 5.000$$

Ingresos anuales	Frecuencia
[10.000 – 15.000]	7
[15.000 – 20.000]	23
[20.000 – 25.000]	4
[25.000 – 30.000]	3
[30.000 – 35.000]	3

Para simplificar los cálculos expresaremos los datos en miles de euros:

Ingresos anuales	n_i
[10 – 15]	7
[15 – 20]	23
[20 – 25]	4
[25 – 30]	3
[30 – 35]	3

a) Media aritmética:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{760}{40} = 19$$

Ingresos anuales	n_i	x_i	$x_i n_i$
[10 – 15]	7	12,5	87,5
[15 – 20]	23	17,5	402,5
[20 – 25]	4	22,5	90

[25 – 30]	3	27,5	82,5
[30 – 35]	3	32,5	97,5
Sumas	40		760

b) Mediana:

El intervalo que contiene a la media será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($40/2 = 20$). Como vemos, la mediana está en el intervalo $[15 – 20]$.

$$Me = L_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i} = 15 + 5 \frac{20 - 7}{23} = 17,83$$

Ingresos anuales	n_i	x_i	$x_i n_i$	N_i
[10 – 15]	7	12,5	87,5	7
[15 – 20]	23	17,5	402,5	30
[20 – 25]	4	22,5	90	34
[25 – 30]	3	27,5	82,5	37
[30 – 35]	3	32,5	97,5	40
Sumas	40		760	

c) Moda:

El intervalo que contiene a la moda será el de mayor frecuencia absoluta. En nuestro caso, se trata del intervalo $[15 – 20]$.

$$Mo = L_{i-1} + a_i \frac{n_{i+1}}{n_{i-1} + n_{i+1}} = 15 + 5 \frac{4}{7 + 4} = 16,82$$

d) Media cuadrática:

$$C = \sqrt{\sum X_i^2 \cdot n_i / N}$$

$$C = \sqrt{15600 / 40} = 19,75$$

Ingresos anuales	n_i	x_i	$x_i n_i$	N_i	$X_i^2 n_i$
[10 – 15]	7	12,5	87,5	7	1093,75
[15 – 20]	23	17,5	402,5	30	7043,75
[20 – 25]	4	22,5	90	34	2025
[25 – 30]	3	27,5	82,5	37	2268,75
[30 – 35]	3	32,5	97,5	40	3168,75
Sumas	40		760		15.600

e) Varianza:

$$S_2 = \frac{\sum (X_i - \bar{X})^2 \cdot n_i}{N - 1} = \frac{1160}{39} = 29,74$$

Ingresos anuales	n_i	x_i	$x_i n_i$	N_i	$(X_i - \bar{X})^2 \cdot n_i$
[10 – 15]	7	12,5	87,5	7	295,75
[15 – 20]	23	17,5	402,5	30	51,75
[20 – 25]	4	22,5	90	34	49
[25 – 30]	3	27,5	82,5	37	216,75
[30 – 35]	3	32,5	97,5	40	546,75
Sumas	40		760		1.160

f) Desviación típica:

Es la raíz cuadrada de la varianza.

$$S = \sqrt{S^2} = \sqrt{29,74} = 5,45$$

g) Coeficiente de variación de Pearson:

$$V = \frac{S}{\bar{X}} * 100 = \frac{5,45}{19} \times 100 = 28,68\%$$

h) Desviación media:

$$DM = \frac{\sum |\bar{X}_i - \bar{X}| \cdot n_i}{N} = \frac{160}{40} = 4$$

Ingresos anuales	n_i	x_i	$x_i n_i$	N_i	$X_i^2 n_i$	$ X_i - \bar{X} n_i$
[10 – 15]	7	12,5	87,5	7	1093,75	45,5
[15 – 20]	23	17,5	402,5	30	7043,75	34,5
[20 – 25]	4	22,5	90	34	2025	14
[25 – 30]	3	27,5	82,5	37	2268,75	25,5
[30 – 35]	3	32,5	97,5	40	3168,75	40,5
Sumas	40		760		15.600	160

i) Desviación mediana:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{136}{40} = 3,42$$

Ingresos anuales	n_i	x_i	$x_i n_i$	N_i	$X_i^2 n_i$	$ X_i - \bar{X} n_i$	$ X_i - Me n_i$
[10 – 15]	7	12,5	87,5	7	1093,75	45,5	37,31
[15 – 20]	23	17,5	402,5	30	7043,75	34,5	7,59
[20 – 25]	4	22,5	90	34	2025	14	18,68
[25 – 30]	3	27,5	82,5	37	2268,75	25,5	29,01
[30 – 35]	3	32,5	97,5	40	3168,75	40,5	44,01
Sumas	40		760		15.600	160	136,6

j) Índice de Gini:

$$G = 1 - \frac{\sum Q_i}{\sum P_i} = 1 - \frac{339,46}{370} = 0,083$$

Ingresos anuales	n_i	x_i	$x_i n_i$	N_i	P_i	Q_i
[10 – 15]	7	12,5	87,5	7	17,5	11,51
[15 – 20]	23	17,5	402,5	30	75	64,47
[20 – 25]	4	22,5	90	34	85	76,31
[25 – 30]	3	27,5	82,5	37	92,5	87,17
[30 – 35]	3	32,5	97,5	40	100	100
Sumas	40		760		370	339,46

P_i = frecuencia relativa acumulada, multiplicada por 100.

Q_i = cada producto $X_i n_i$, haciéndolo relativo y acumulado, multiplicado por 100.

k) Coeficiente de asimetría de Pearson:

$$A = \frac{\bar{X} - M_o}{S} = \frac{19 - 16,82}{5,45} = 0,40$$

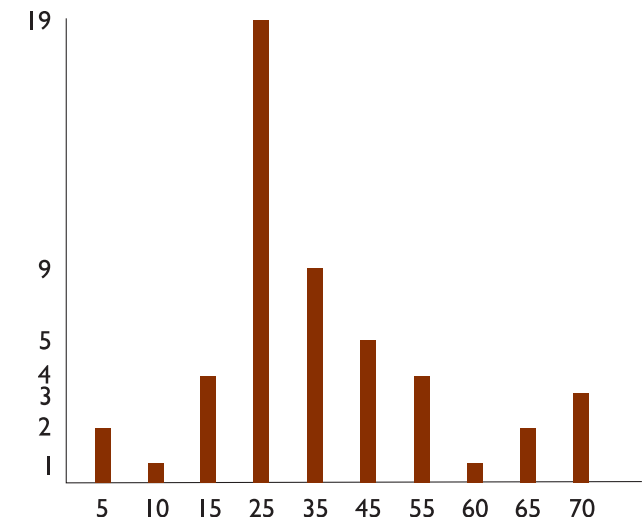
Ejercicio 4

l) Tabla tipo II

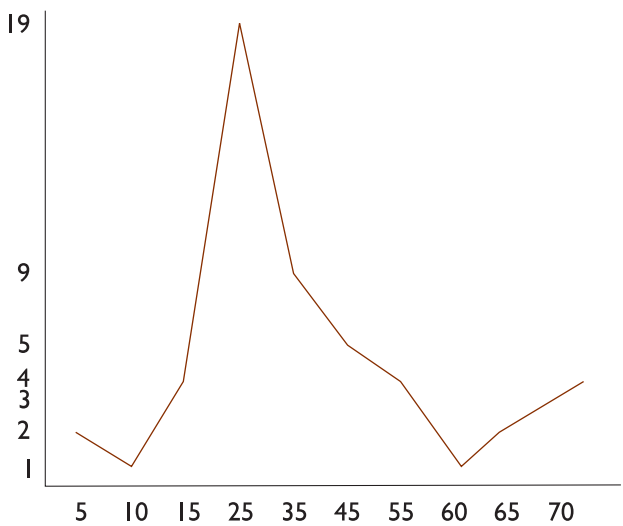
La tabla de frecuencias quedaría como sigue:

Edad	Frecuencia
5	2
10	1
15	4
25	19
35	9
45	5
55	4
60	1
65	2
70	3

a) Diagrama de barras:



b) Polígono de frecuencias:



c) ¿Qué porcentaje de espectadores supone cada edad? Para ello, calcularemos las frecuencias relativas y posteriormente las multiplicaremos por 100 para obtener el porcentaje.

x_i	n_i	f_i	%
5	2	0,04	4%
10	1	0,02	2%
15	4	0,08	8%
25	19	0,38	38%
35	9	0,18	18%
45	5	0,10	10%
55	4	0,08	8%
60	1	0,02	2%
65	2	0,04	4%
70	3	0,06	6%
Sumas	50	1	100%

d) Media aritmética:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{1715}{50} = 34,3$$

x_i	n_i	f_i	%	$x_i n_i$
5	2	0,04	4%	10
10	1	0,02	2%	10
15	4	0,08	8%	60
25	19	0,38	38%	475
35	9	0,18	18%	315
45	5	0,10	10%	225
55	4	0,08	8%	220
60	1	0,02	2%	60
65	2	0,04	4%	130
70	3	0,06	6%	210
Sumas	50	1	100%	1715

e) Mediana:

La mediana será el primer valor cuya frecuencia absoluta acumulada supere a $N/2$, o sea, el primer x_i cuya N_i supere $50/2 = 25$.

x_i	n_i	f_i	%	$x_i n_i$	N_i
5	2	0,04	4%	10	2
10	1	0,02	2%	10	3
15	4	0,08	8%	60	7
25	19	0,38	38%	475	26
35	9	0,18	18%	315	35
45	5	0,10	10%	225	40
55	4	0,08	8%	220	44
60	1	0,02	2%	60	45
65	2	0,04	4%	130	47
70	3	0,06	6%	210	50
Sumas	50	1	100%	1715	

Como vemos, la mediana será 25, que es el primer valor de la variable cuya frecuencia absoluta acumulada (26) supera a $N/2$ (25).

f) Moda:

La moda será aquel valor de la variable con mayor frecuencia absoluta. En este caso, la moda es 25.

II) Tabla tipo III

En primer lugar, hemos de calcular el rango, que es la diferencia entre el mayor valor y el menor:

$$\text{Rango} = 70 - 5 = 65$$

Como queremos determinar 7 clases de la misma amplitud, dividiremos el rango entre el número de clases para calcular esa amplitud:

$$\text{Amplitud} = \text{Rango} / N^{\circ} \text{ clases} = 65 / 7 = 9,29$$

Para facilitar la construcción de la tabla redondearemos esa amplitud, siempre al alza para poder cubrir todo el rango. La amplitud será, entonces, de 10.

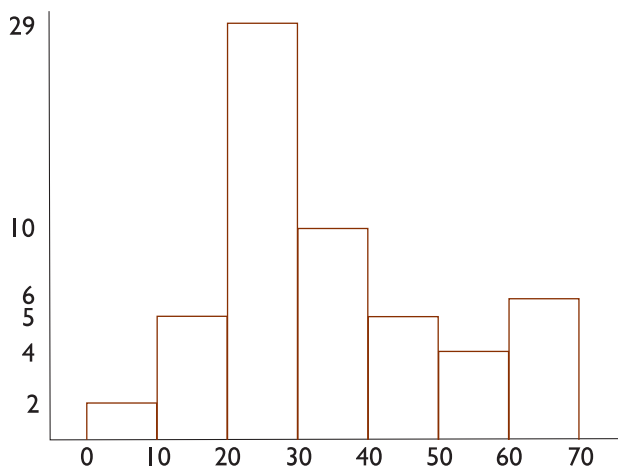
En ese caso, como habrán 7 intervalos de amplitud 10, el rango que cubriremos será de 70 (7×10), lo cual implica que nos excederemos en 5 (ya que el rango original es de 65).

Podemos decidir comenzar con una edad de 5 años antes (0), acabar 5 años después (75), o bien cualquier otra solución intermedia, siempre procurando facilitar los cálculos sin restar información. Para este caso, podemos empezar en 0 años.

La tabla de frecuencias quedaría como sigue:

Edad	Frecuencia
[0 – 10]	2
[10 – 20]	5
[20 – 30]	19
[30 – 40]	9
[40 – 50]	5
[50 – 60]	4
[60 – 70]	6

a) Histograma:



b) Media aritmética:

$$\bar{X} = \frac{\sum x_i \cdot n_i}{N} = \frac{1710}{50} = 34,2$$

Edad	n_i	x_i	$x_i n_i$
[0 – 10]	2	5	10
[10 – 20]	5	15	75
[20 – 30]	19	25	475
[30 – 40]	9	35	315
[40 – 50]	5	45	225
[50 – 60]	4	55	220
[60 – 70]	6	65	390
Sumas	50		1710

c) Mediana:

El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($50/2 = 25$). Como vemos, la mediana está en el intervalo [20 – 30].

$$Me = L_{i-1} + a_i \frac{N/2 - n_{i-1}}{n_i} = 20 + 10 \frac{25 - 7}{19} = 29,47$$

Ingresos anuales	n_i	x_i	$x_i n_i$	N_i
[0 – 10]	2	5	10	2
[10 – 20]	5	15	75	7
[20 – 30]	19	25	475	26
[30 – 40]	9	35	315	35
[40 – 50]	5	45	225	40
[50 – 60]	4	55	220	44
[60 – 70]	6	65	390	50
Sumas	50		1710	

d) Moda:

El intervalo que contiene a la moda será el de mayor frecuencia absoluta. En nuestro caso, se trata del intervalo [20 – 30].

$$Mo = L_{i-1} + a_i \frac{n_{i+1}}{n_{i-1} + n_{i+1}} = 20 + 10 \frac{9}{5 + 9} = 26,43$$

Ejercicio 5

1) Intervalos regulares

En primer lugar, hemos de calcular el rango, que es la diferencia entre el mayor valor y el menor:

$$\text{Rango} = 50 - 0 = 50$$

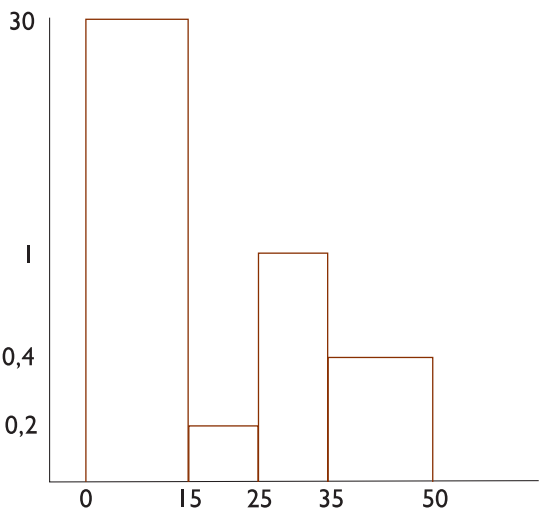
Como queremos determinar 5 clases de la misma amplitud, dividiremos el rango entre el número de clases para calcular esa amplitud:

$$\text{Amplitud} = \text{Rango} / \text{N}^\circ \text{ clases} = 50 / 5 = 10$$

La tabla de frecuencias quedaría como sigue:

Nº veces	Frecuencia
[0 – 10]	30
[10 – 20]	4
[20 – 30]	10
[30 – 40]	4
[40 – 50]	2

a) Histograma:



b) Media aritmética:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{690}{50} = 13,8$$

Nº veces	n_i	x_i	$x_i n_i$
[0 – 10]	30	5	150
[10 – 20]	4	15	60
[20 – 30]	10	25	250
[30 – 40]	4	35	140
[40 – 50]	2	45	90
Sumas	50		690

c) Mediana:

El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($50/2 = 25$). Como vemos, la mediana está en el intervalo [0 – 10].

$$Me = L_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i} = 0 + 10 \frac{25 - 0}{30} = 8,33$$

Nº veces	n_i	x_i	$x_i n_i$	N_i
[0 – 10]	30	5	150	30
[10 – 20]	4	15	60	34
[20 – 30]	10	25	250	44
[30 – 40]	4	35	140	48
[40 – 50]	2	45	90	50
Sumas	50		1710	

d) Moda:

El intervalo que contiene a la moda será el de mayor frecuencia absoluta. En nuestro caso, se trata del intervalo [0 – 10]. Como es el primer intervalo, tenemos que aplicar la fórmula específica:

$$Mo = L_{i-1} + a_i \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

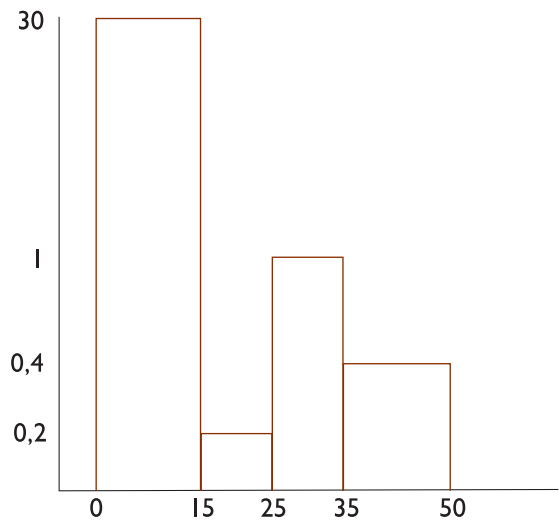
$$Mo = 0 + 10 \frac{30 - 0}{(30 - 0) + (30 - 4)} = 5,36$$

II) Intervalos irregulares

La tabla de frecuencias quedaría como sigue:

Nº veces	Frecuencia
[0 – 15]	32
[15 – 25]	2
[25 – 35]	10
[35 – 50]	6

a) Histograma:



Para confeccionarlo, como se trata de intervalos de diferente amplitud, hemos de calcular las alturas (h_i):

Nº veces	n_i	h_i
[0 – 15]	32	2,13
[15 – 25]	2	0,2
[25 – 35]	10	1
[35 – 50]	6	0,4

b) ¿Qué porcentaje dentro del total supone cada intervalo?

Para ello hemos de calcular la frecuencia relativa y posteriormente multiplicarla por 100 para obtener el porcentaje:

Nº veces	n_i	h_i	f_i	%
[0 – 15]	32	2,13	0,64	64%
[15 – 25]	2	0,2	0,04	4%
[25 – 35]	10	1	0,2	20%
[35 – 50]	6	0,4	0,12	12%
Sumas	50		1	100%

c) Media aritmética:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{835}{50} = 16,7$$

Nº veces	n_i	h_i	x_i	$x_i n_i$
[0 – 15]	32	2,13	7,5	240
[15 – 25]	2	0,2	20	40
[25 – 35]	10	1	30	300
[35 – 50]	6	0,4	42,5	255
Sumas	50			835

d) Mediana:

El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($50/2 = 25$). Como vemos, la mediana está en el intervalo [0 – 15].

$$Me = L_{i-1} + a_i \frac{N/2 - n_{i-1}}{n_i} = 0 + 10 \frac{25 - 0}{32} = 7,81$$

Nº veces	n_i	h_i	x_i	$x_i n_i$	N_i
[0 – 15]	32	2,13	7,5	240	32
[15 – 25]	2	0,2	20	40	34
[25 – 35]	10	1	30	300	44
[35 – 50]	6	0,4	42,5	255	50
Sumas	50			835	

e) Moda:

El intervalo que contiene a la moda será el de mayor altura (h_i). En nuestro caso, se trata del intervalo [0 – 15]. Como es el primer intervalo y además son intervalos de diferente amplitud, tenemos que aplicar la fórmula específica para este caso:

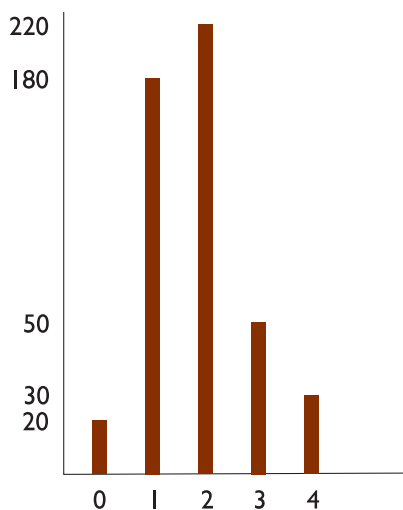
$$Mo = L_{i-1} + a_i \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})}$$

$$Mo = 0 + 10 \frac{2,13 - 0}{(2,13 - 0) + (2,13 - 0,2)} = 5,02$$

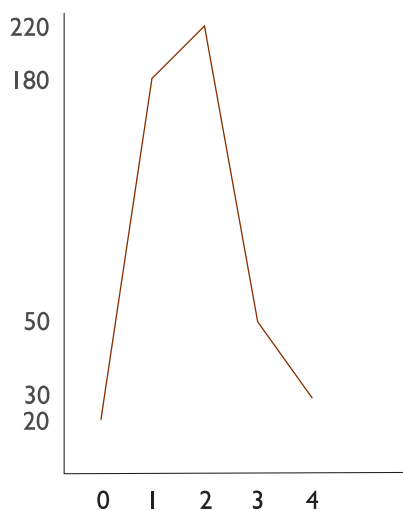
Ejercicio 6

a) Diagrama de barras:

X_i	n_i
0	20
1	180
2	220
3	50
4	30



b) Polígono de frecuencias:



c) Número de coches más habitual en las familias. Se trata de calcular la moda. La moda, en este tipo de distribuciones, será aquel valor de X_i con mayor frecuencia absoluta.

En este caso, la moda es 2.

d) ¿Es representativa la media aritmética? Para saberlo hemos de calcular el coeficiente de variación de Pearson. Dado que se trata de una encuesta a la población total, calcularemos la media y varianza poblacional.

$$V = \frac{\sigma}{\mu} * 100$$

Calcularemos los datos necesarios.

Media aritmética:

$$\mu = \frac{\sum x_i n_i}{N} = \frac{890}{500} = 1,78$$

X_i	n_i	$x_i n_i$
0	20	0
1	180	180
2	220	440
3	50	150
4	30	120
Sumas	500	890

Desviación típica:

Es la raíz cuadrada de la varianza. Calculamos primero la varianza.

$$\sigma = \frac{\sum X_i^2 \cdot n_i}{N} - \mu^2 = \frac{1990}{500} - 1,78^2 = 0,8116$$

X_i	n_i	$x_i n_i$	$X_i^2 n_i$
0	20	0	0
1	180	180	180
2	220	440	880
3	50	150	450
4	30	120	480
Sumas	500	890	1.990

$$\sigma = \sqrt{\sigma^2} = \sqrt{0,8116} = 0,90$$

Por tanto, el coeficiente de variación será el siguiente:

$$V = \frac{\sigma}{\mu} * 100$$

$$V = \frac{0,90}{1,78} * 100 = 50,56\%$$

La media aritmética es poco representativa, ya que se encuentra entre $40\% < V < 60\%$.

e) ¿Sería más representativa la mediana?

Para responder a esta cuestión debemos calcular la desviación media y compararla con la desviación mediana. Si esta última fuera inferior, sería más representativa la mediana.

Desviación media:

$$DM = \frac{\sum |\bar{X}_i - \mu| \cdot n_i}{N} = \frac{352}{500} = 0,704$$

X_i	n_i	$x_i n_i$	$X_i^2 n_i$	$ X_i - \bar{X} n_i$	N_i
0	20	0	0	35,6	20
1	180	180	180	140,4	200
2	220	440	880	48,4	420
3	50	150	450	61	470
4	30	120	480	66,6	500
Sumas	500	890	1.990	352	

Desviación mediana:

Calcularemos primero la mediana. La mediana será el primer valor de X_i cuya frecuencia absoluta acumulada supere a $N/2$ ($500/2 = 250$). Como vemos, la mediana es 2.

La desviación mediana la calcularemos con la siguiente fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{330}{500} = 0,66$$

X_i	n_i	$x_i n_i$	$X_i^2 \cdot n_i$	$ X_i - \bar{X} \cdot n_i$	N_i	$ X_i - Me \cdot n_i$
0	20	0	0	35,6	20	40
1	180	180	180	140,4	200	180
2	220	440	880	48,4	420	0
3	50	150	450	61	470	50
4	30	120	480	66,6	500	60
Sumas	500	890	1.990	352		330

Como vemos, la desviación mediana es menor que la desviación media. Por ello, podemos decir que la mediana es más representativa de esta distribución que la media aritmética.

f) El porcentaje de familias con 2 coches será la frecuencia relativa multiplicada por 100, o sea 44% ($220 \times 100/500$).

Ejercicio 7

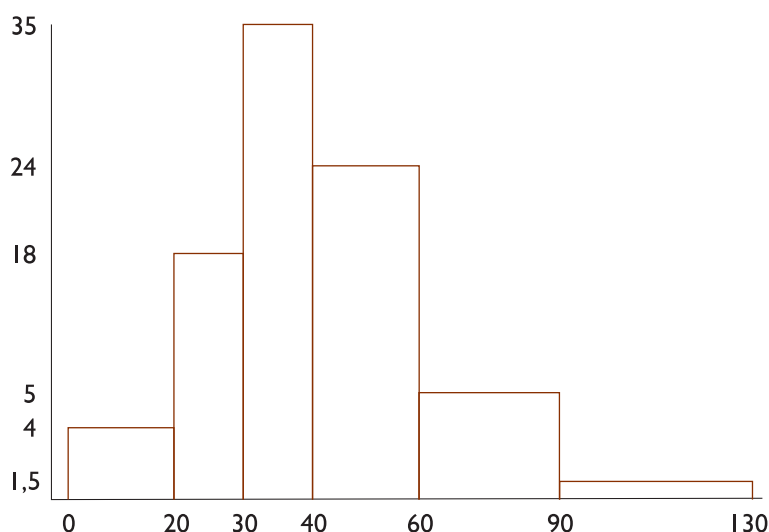
La tabla de frecuencias quedaría como sigue:

Latas de cerveza al mes	n_i
[0 – 20]	80
[20 – 30]	180
[30 – 40]	350
[40 – 60]	480
[60 – 90]	150
[90 – 130]	60

a) Histograma:

Como se trata de intervalos de diferente amplitud hemos de representar el histograma con las alturas (h_i):

Clases	n_i	h_i
[0 – 20]	80	4
[20 – 30]	180	18
[30 – 40]	350	35
[40 – 60]	480	24
[60 – 90]	150	5
[90 – 130]	60	1,5



b) ¿Cuál es el consumo de cerveza al mes más habitual en las familias? Se trata de calcular la moda.

El intervalo que contiene a la moda será el de mayor altura (h_i). En nuestro caso, se trata del intervalo [30 – 40]. Tenemos que aplicar la fórmula específica para este caso:

c) ¿Es representativa la media aritmética? Para saberlo hemos de calcular el coeficiente de variación de Pearson. Dado que es una muestra, calcularemos la media y varianza muestral.

$$Mo = L_{i-1} + a_i \frac{h_{i+1}}{h_{i-1} + h_{i+1}}$$

$$M_o = 30 + 10 \frac{24}{18 + 24} = 35,71$$

Media aritmética:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{59.400}{1300} = 45,69$$

Clases	n_i	x_i	$x_i n_i$
[0 – 20]	80	10	800
[20 – 30]	180	25	4500
[30 – 40]	350	35	12250
[40 – 60]	480	50	24000
[60 – 90]	150	75	11250
[90 – 130]	60	110	6600
Sumas	1.300		59.400

Desviación típica:

Es la raíz cuadrada de la varianza. Calculamos primero la varianza:

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{N-1} = \frac{604876,93}{1299} = 465,65$$

Clases	n_i	x_i	$x_i n_i$	$(X_i - \bar{X})^2 \cdot n_i$
[0 – 20]	80	10	800	101902,088
[20 – 30]	180	25	4500	77053,698
[30 – 40]	350	35	12250	39996,635
[40 – 60]	480	50	24000	8916,528
[60 – 90]	150	75	11250	128861,415
[90 – 130]	60	110	6600	248146,566
Sumas	1.300		59.400	604.876,93

$$S = \sqrt{s^2} = \sqrt{465,65} = 21,58$$

Por tanto, el coeficiente de variación será el siguiente:

$$V = \frac{S}{\bar{X}} * 100 =$$

$$V = \frac{21,58}{45,69} * 100 = 47,23\%$$

La media aritmética es poco representativa, ya que se encuentra entre $40\% < V < 60\%$.

d) ¿Sería más representativa la mediana?

Para responder a esta cuestión debemos calcular la desviación media y compararla con la desviación mediana. Si esta última fuera inferior, sería más representativa la mediana.

Desviación media

$$DM = \frac{\sum |X_i - \bar{X}| \cdot n_i}{N} = \frac{20644,8}{1300} = 15,88$$

Clases	n_i	x_i	$x_i n_i$	$X_i^2 n_i$	$ X_i - \bar{X} n_i$	N_i
[0 – 20]	80	10	800	8000	2855,2	80
[20 – 30]	180	25	4500	112500	3724,2	260
[30 – 40]	350	35	12250	428750	3741,5	610
[40 – 60]	480	50	24000	1200000	2068,8	1090
[60 – 90]	150	75	11250	843750	4396,5	1240
[90 – 130]	60	110	6600	726000	3858,6	1300
Sumas	1.300		59.400	3.319.000	20644,8	

Desviación mediana:

Calcularemos primero la mediana. El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($1300/2 = 650$). Como vemos, la mediana está en el intervalo [40 – 60].

$$Me = L_{i-1} + a_i \frac{N/2 - n_{i-1}}{n_i} = 40 + 20 \frac{650 - 610}{480} = 41,67$$

La desviación mediana la calcularemos con la siguiente fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{20965,6}{1300} = 16,13$$

x_i	n_i	$x_i n_i$	$X^2_i n_i$	$ X_i - \bar{X} n_i$	N_i	$ X_i - Me n_i$
10	80	800	8000	2855,2	80	2532,8
25	180	4500	112500	3724,2	260	3000,6
35	350	12250	428750	3741,5	610	2334,5
50	480	24000	1200000	2068,8	1090	3998,4
75	150	11250	843750	4396,5	1240	4999,5
110	60	6600	726000	3858,6	1300	4099,8
	1.300	59.400	3.319.000	20644,8		20.965,6

Como vemos, la desviación mediana es mayor que la desviación media. Por ello, podemos decir que la media aritmética es más representativa de esta distribución que la mediana.

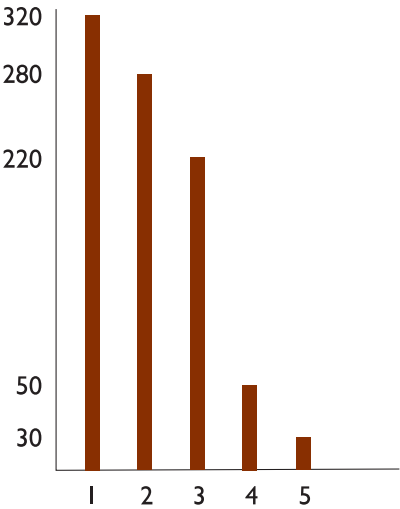
e) El porcentaje de familias será la frecuencia relativa multiplicada por 100.

Clases	n_i	f_i	%
[0 – 20]	80	0,0615	6,15%
[20 – 30]	180	0,1385	13,85%
[30 – 40]	350	0,2692	26,92%
[40 – 60]	480	0,3692	36,92%
[60 – 90]	150	0,1154	11,54%
[90 – 130]	60	0,0462	4,62%
Sumas	1.300	1	100%

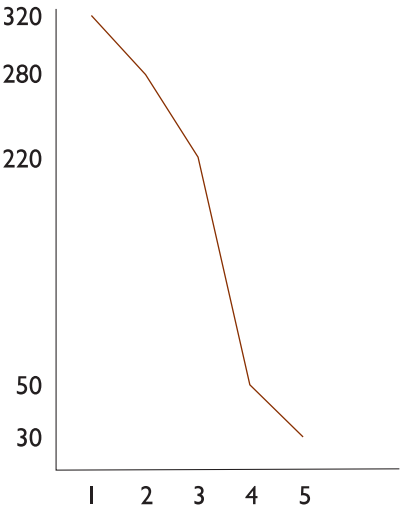
Ejercicio 8

a) Diagrama de barras:

X_i	n_i
1	320
2	280
3	220
4	50
5	30



b) Polígono de frecuencias:



c) Número de componentes más habitual en las familias. Se trata de calcular la moda. La moda, en este tipo de distribuciones, será aquel valor de X_i con mayor frecuencia absoluta.

En este caso, la moda es 1.

d) ¿Es representativa la media aritmética? Para saberlo hemos de calcular el coeficiente de variación de Pearson. Al tratarse de una muestra, calcularemos la media y varianza muestral.

$$V = \frac{S}{\bar{X}} = *100$$

Calcularemos los datos necesarios.

Media aritmética:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{1890}{900} = 2,1$$

X_i	n_i	$x_i n_i$
1	320	320
2	280	560
3	220	660
4	50	200
5	30	150
Sumas	900	1.890

Desviación típica:

Es la raíz cuadrada de la varianza. Calculamos primero la varianza.

$$S^2 = \frac{\sum (X_i - \bar{X})^2 \cdot n_i}{N - 1} = \frac{1001}{899} = 1,11346$$

X_i	n_i	$x_i n_i$	$(X_i - \bar{X})^2 \cdot n_i$
1	320	320	387,2
2	280	580	2,8

3	220	660	178,2
4	50	200	180,5
5	30	150	252,3
Sumas	900	1.910	1.001

$$S = \sqrt{S^2} = \sqrt{1,11346} = 1,05521$$

Por tanto, el coeficiente de variación será el siguiente:

$$V = \frac{S}{\bar{X}} * 100$$

$$V = \frac{1,05}{2,1} * 100 = 50\%$$

La media aritmética es poco representativa, ya que se encuentra entre $40\% < V < 60\%$.

e) ¿Sería más representativa la mediana?

Para responder a esta cuestión debemos calcular la desviación media y compararla con la desviación mediana. Si esta última fuera inferior, sería más representativa la mediana.

Desviación media:

$$DM = \frac{\sum |X_i - \bar{X}| \cdot n_i}{N} = \frac{1543,6}{900} = 1,72$$

X_i	n_i	$x_i n_i$	$X_i^2 n_i$	$ X_i - \bar{X} n_i$	N_i
1	320	320	320	358,4	320
2	280	560	1120	33,6	600
3	220	660	1980	193,6	820
4	50	200	800	94	870
5	30	150	750	864	900
Sumas	900	1.890	4.970	1.543,6	

Desviación mediana:

Calcularemos primero la mediana. La mediana será el primer valor de X_i cuya frecuencia absoluta acumulada supere a $N/2$ ($900/2 = 450$). Como vemos, la mediana es 2.

La desviación mediana la calcularemos con la siguiente fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{730}{900} = 0,81$$

X_i	n_i	$x_i n_i$	$X_i^2 n_i$	$ X_i - \bar{X} n_i$	N_i	$ X_i - Me n_i$
1	320	320	320	358,4	320	320
2	280	560	1120	33,6	600	0
3	220	660	1980	193,6	820	220
4	50	200	800	94	870	100
5	30	150	750	864	900	90
Sumas	900	1.890	4.970	1.543,6		730

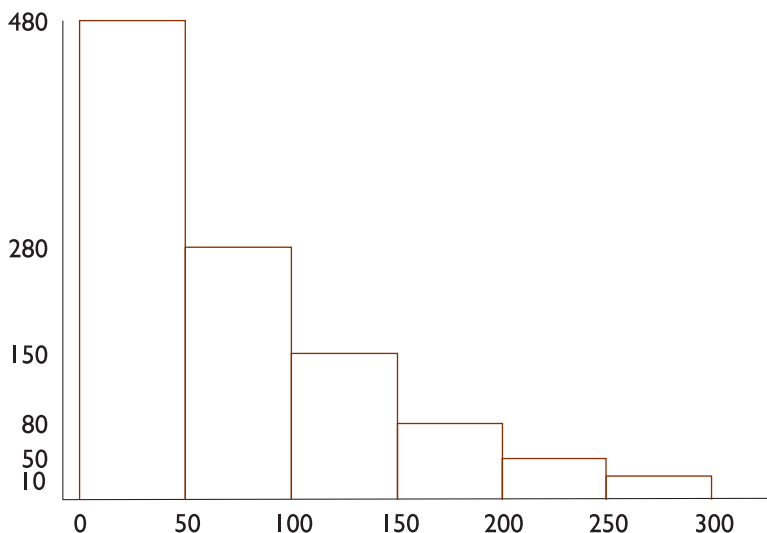
Como vemos, la desviación mediana es menor que la desviación media. Por ello, podemos decir que la mediana es más representativa de esta distribución que la media aritmética.

Ejercicio 9

La tabla de frecuencias quedaría como sigue:

Consumo eléctrico €	n_i
[0 – 50]	480
[50 – 100]	280
[100 – 150]	150
[150 – 200]	80
[200 – 250]	50
[250 – 300]	10

a) Histograma:



b) ¿Cuál es el consumo eléctrico al mes más habitual? Se trata de calcular la moda.

El intervalo que contiene a la moda será el de mayor frecuencia absoluta (n_i). En nuestro caso, se trata del intervalo $[0 - 50]$. Como es el primer intervalo, debemos aplicar la fórmula específica para este caso:

$$Mo = L_{i-1} + a_i \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

$$Mo = 0 + 50 \frac{480}{(480 - 0) + (480 - 280)} = 35,29$$

c) ¿Es representativa la media aritmética? Para saberlo hemos de calcular el coeficiente de variación de Pearson. Dado que es un estudio poblacional, calcularemos la media y varianza poblacional.

$$V = \frac{\sigma}{\mu} * 100$$

Calcularemos los datos necesarios.

Media aritmética:

$$\mu = \frac{\sum x_i n_i}{N} = \frac{79750}{1050} = 75,95$$

Clases	n_i	x_i	$x_i n_i$
[0 – 50]	480	25	12000
[50 – 100]	280	75	21000
[100 – 150]	150	125	18750
[150 – 200]	80	175	14000
[200 – 250]	50	225	11250
[250 – 300]	10	275	2750
Sumas	1.050		79.750

Desviación típica:

Es la raíz cuadrada de la varianza. Calculamos primero la varianza:

$$\sigma^2 = \frac{\sum X_i^2 \cdot n_i}{N} - \mu^2 = \frac{9956250}{1050} - 75,95^2 = 3.714,81$$

Clases	n_i	x_i	$x_i n_i$	$X_i^2 n_i$
[0 – 50]	480	25	12000	300000
[50 – 100]	280	75	21000	1575000
[100 – 150]	150	125	18750	2343750
[150 – 200]	80	175	14000	2450000
[200 – 250]	50	225	11250	2531250
[250 – 300]	10	275	2750	756250
Sumas	1.050		79.750	9.956.250

$$\sigma = \sqrt{\sigma^2} = \sqrt{3714,81} = 60,95$$

Por tanto, el coeficiente de variación será el siguiente:

$$V = \frac{\sigma}{\mu} * 100$$

$$V = \frac{60,95}{75,95} * 100 = 80,25\%$$

La media aritmética no es nada representativa, ya que es superior al 60%.

d) ¿Sería más representativa la mediana?

Para responder a esta cuestión debemos calcular la desviación media y compararla con la desviación mediana. Si esta última fuera inferior, sería más representativa la mediana.

Desviación media:

$$DM = \frac{\sum |X_i - \mu| \cdot n_i}{N} = \frac{49446}{1050} = 47,09$$

Clases	n_i	x_i	$x_i n_i$	$ X_i - \bar{X} n_i$	N_i
[0 – 50]	480	25	12000	24456	480
[50 – 100]	280	75	21000	266	760
[100 – 150]	150	125	18750	7357,5	910
[150 – 200]	80	175	14000	7924	990
[200 – 250]	50	225	11250	7452,5	1040
[250 – 300]	10	275	2750	1990,5	1050
Sumas	1.050		79.750	49.446,5	

Desviación mediana:

Calcularemos primero la mediana. El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($1050/2 = 525$). Como vemos, la mediana está en el intervalo [50 – 100].

$$Me = L_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i} = 50 + 50 \frac{525 - 480}{280} = 58,04$$

La desviación mediana la calcularemos con la siguiente fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{50526,4}{1050} = 48,12$$

Clases	n_i	x_i	$x_i n_i$	$ X_i - \bar{X} n_i$	$ X_i - Me n_i$
[0 – 50]	480	25	12000	24456	15859,2
[50 – 100]	280	75	21000	266	4748,8
[100 – 150]	150	125	18750	7357,5	10044
[150 – 200]	80	175	14000	7924	9356,8
[200 – 250]	50	225	11250	7452,5	8348
[250 – 300]	10	275	2750	1990,5	2169,6
Sumas	1.050		79.750	49.446,5	50.526,4

Como vemos, la desviación mediana es mayor que la desviación media. Por ello, podemos decir que la media aritmética es más representativa de esta distribución que la mediana.

Ejercicio 10

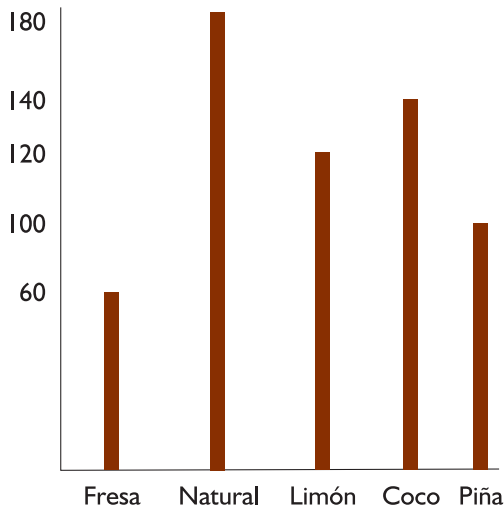
a) En una estadística de atributos la única medida de tendencia central que puede obtenerse es la moda.

La moda corresponderá al atributo que mayor frecuencia absoluta presente. En este caso, la moda (lo más habitual) es el sabor NATURAL.

b) Para calcular el porcentaje de encuestados que prefiere cada sabor hemos de determinar la frecuencia relativa y posteriormente multiplicarla por 100 para obtener el porcentaje.

Sabor	n_i	f_i	%
Fresa	60	0,10	10%
Natural	180	0,30	30%
Limón	120	0,20	20%
Coco	140	0,233	23,3%
Piña	100	0,167	16,7%
Sumas	600	1	100%

c) Diagrama de barras:



Ejercicio I I

a) Tabla de tipo II:

Viviendas	n_i
1	16
2	8
3	3
4	2
5	1
Sumas	30

b) ¿Sería más representativa la media aritmética o la mediana?

Para responder a esta cuestión debemos calcular la desviación media y compararla con la desviación mediana. Si esta última fuera inferior, sería más representativa la mediana.

Desviación media:

En primer lugar calcularemos la media aritmética.

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{54}{30} = 1,8$$

X_i	n_i	$x_i n_i$
1	16	16
2	8	16
3	3	9
4	2	8
5	1	5
Sumas	30	54

La desviación media será la siguiente:

$$DM = \frac{\sum |X_i - \bar{X}| \cdot n_i}{N} = \frac{25,6}{30} = 0,85$$

X_i	n_i	$x_i n_i$	$ X_i - \bar{X} \cdot n_i$	N_i
1	16	16	12,8	16
2	8	16	1,6	24
3	3	9	3,6	27
4	2	8	4,4	29
5	1	5	3,2	30
Sumas	30	54	25,6	

Desviación mediana:

Calcularemos primero la mediana. La mediana será el primer valor de X_i cuya frecuencia absoluta acumulada supere a $N/2$ ($30/2 = 15$). Como vemos, la mediana es 1.

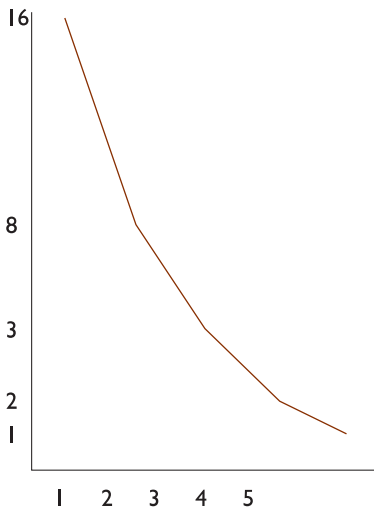
La desviación mediana la calcularemos con la siguiente fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{24}{30} = 0,80$$

X_i	n_i	$x_i n_i$	$ X_i - \check{X} n_i$	N_i	$ X_i - Me n_i$
1	16	16	12,8	16	0
2	8	16	1,6	24	8
3	3	9	3,6	27	6
4	2	8	4,4	29	6
5	1	5	3,2	30	4
Sumas	30	54	25,6		24

Como vemos, la desviación mediana es menor que la desviación media. Por ello, podemos decir que la mediana es más representativa de esta distribución que la media aritmética.

c) Polígono de frecuencias:



Ejercicio 12

a) ¿Es representativa la media aritmética? Para saberlo hemos de calcular el coeficiente de variación de Pearson. Dado que se trata de una muestra, calcularemos la media y varianza muestral.

$$V = \frac{S}{\bar{X}} * 100$$

Calcularemos los datos necesarios.

Media aritmética:

$$\bar{X} = \frac{\sum x_i \cdot n_i}{N} = \frac{19800}{200} = 99$$

Clases	n_i	x_i	$x_i n_i$
[0 – 50]	10	25	250
[50 – 80]	40	65	2600
[80 – 100]	80	90	7200
[100 – 130]	30	115	3450
[130 – 150]	20	140	2800
[150 – 200]	20	175	3500
Sumas	200		19.800

Desviación típica

Es la raíz cuadrada de la varianza. Calculamos primero la varianza:

$$S_2 = \frac{\sum (X_i - \bar{X})^2 \cdot n_i}{N - 1} = \frac{26430}{199} = 1.328,14$$

Clases	n_i	x_i	$x_i n_i$	$(X_i - \bar{X})^2 n_i$
[0 – 50]	10	25	250	54760
[50 – 80]	40	65	2600	46240
[80 – 100]	80	90	7200	6480
[100 – 130]	30	115	3450	7680
[130 – 150]	20	140	2800	33620
[150 – 200]	20	175	3500	115520
Sumas	200		19.800	264.300

$$S = \sqrt{S^2} = \sqrt{1328,14} = 36,44$$

Por tanto, el coeficiente de variación será el siguiente:

$$V = \frac{S}{\overline{X}} * 100$$

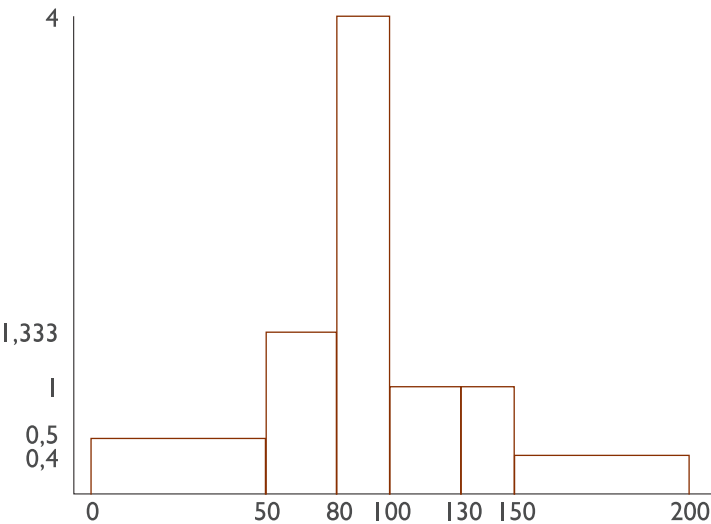
$$V = \frac{36,44}{99} \times 100 = 36,81\%$$

La media aritmética es representativa de esta distribución, ya que es inferior al 40%.

b) Histograma:

Como se trata de intervalos de diferente amplitud hemos de representar el histograma con las alturas (h_i):

Clases	n_i	h_i
[0 – 50]	10	0,5
[50 – 80]	40	1,333
[80 – 100]	80	4
[100 – 130]	30	1
[130 – 150]	20	1
[150 – 200]	20	0,4



Ejercicio 13

a) Tabla tipo III:

En primer lugar, hemos de calcular el rango, que es la diferencia entre el mayor valor y el menor:

$$\text{Rango} = 50 - 0 = 50$$

Como queremos determinar 5 clases de la misma amplitud, dividiremos el rango entre el número de clases para calcular esa amplitud:

$$\text{Amplitud} = \text{Rango} / N^{\circ} \text{ clases} = 50 / 5 = 10$$

Libros	Frecuencia
[0 – 10]	20
[10 – 20]	10
[20 – 30]	5
[30 – 40]	3
[40 – 50]	2

b) Concentración:

Calcularemos el índice de Gini.

$$G = 1 - \frac{\sum Q_i}{\sum P_i} = 1 - \frac{311,4}{407,5} = 0,2358$$

Libros	n_i	x_i	$x_i n_i$	f_i	P_i	Q_i
[0 – 10]	20	5	100	0,5	50	17,54
[10 – 20]	10	15	150	0,25	75	43,86
[20 – 30]	5	25	125	0,125	87,5	65,79
[30 – 40]	3	35	105	0,075	95	84,21
[40 – 50]	2	45	90	0,05	100	100
Sumas	40		570	1	407,5	311,4

P_i = frecuencia relativa acumulada, multiplicada por 100

Q_i = cada producto $X_i n_i$, haciéndolo relativo y acumulado, multiplicado por 100.

Existe poca concentración (diríamos que hay mucha concentración cuando $G > 0,75$). Por tanto, la moda, según este estudio, no sería la medida de tendencia central más representativa de esta distribución.

c) Moda:

El intervalo que contiene a la moda será el de mayor frecuencia absoluta. En nuestro caso, se trata del intervalo $[0 - 10]$. Como es el primer intervalo, tenemos que aplicar la fórmula específica:

$$M_o = L_{i-1} + a_i \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

$$M_o = 0 + 10 \frac{20 - 0}{(20 - 0) + (20 - 10)} = 6,67$$

Este dato significa que el número más habitual de libros que leen cada año los individuos de la muestra es de 6,67.

Ejercicio 14

a) Asimetría:

Calcularemos el coeficiente de asimetría de Pearson, simbolizado por A, con la siguiente fórmula:

$$A = \frac{\bar{X} - M_o}{S}$$

Para ello, hemos de calcular previamente la media aritmética, la moda y la desviación típica.

Media aritmética

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{5400}{240} = 22,5$$

Clases	n_i	x_i	$x_i n_i$
$[0 - 10]$	80	5	400

[10 – 30]	90	20	1800
[30 – 40]	40	35	1400
[40 – 60]	20	50	1000
[60 – 100]	10	80	800
Sumas	240		5.400

Moda:

Como se trata de intervalos irregulares hemos de calcular las alturas.

Clases	n_i	x_i	$x_i n_i$	h_i
[0 – 10]	80	5	400	8
[10 – 30]	90	20	1800	4,5
[30 – 40]	40	35	1400	4
[40 – 60]	20	50	1000	1
[60 – 100]	10	80	800	0,25
Sumas	240		5.400	

El intervalo que contiene a la moda será el de mayor altura (h_i). En nuestro caso, se trata del intervalo [0 – 10]. Como es el primer intervalo y además son intervalos de diferente amplitud, tenemos que aplicar la fórmula específica para este caso:

$$Mo = L_{i-1} + a_i \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})}$$

$$Mo = 0 + 10 \frac{8 - 0}{(8 - 0) + (8 - 4,5)} = 6,96$$

Desviación típica

Es la raíz cuadrada de la varianza. Calculamos primero la varianza.

$$S_2 = \frac{\sum (X_i - \bar{X})^2 \cdot n_i}{N - 1} = \frac{79500}{239} = 332,74$$

Clases	n_i	x_i	$x_i n_i$	h_i	$(X_i - \bar{X})^2 n_i$
[0 – 10]	80	5	400	8	24500
[10 – 30]	90	20	1800	4,5	562,5
[30 – 40]	40	35	1400	4	6250
[40 – 60]	20	50	1000	1	15125
[60 – 100]	10	80	800	0,25	33062,5
Sumas	240		5.400		79.500

$$S = \sqrt{S^2} = \sqrt{332,64} = 18,24$$

Con todo ello, el coeficiente de asimetría de Pearson resulta:

$$A = \frac{\bar{X} - Mo}{S} = \frac{22,5 - 6,96}{18,24} = 0,85$$

Como es superior a 0,65 podemos decir que existe mucha asimetría. Por ello, la mediana puede ser la medida de tendencia central más significativa de esta distribución.

b) Mediana:

El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($240/2 = 120$).

Clases	n_i	x_i	$x_i n_i$	h_i	$X_i^2 n_i$	N_i
[0 – 10]	80	5	400	8	2000	80
[10 – 30]	90	20	1800	4,5	36000	170
[30 – 40]	40	35	1400	4	49000	210
[40 – 60]	20	50	1000	1	50000	230
[60 – 100]	10	80	800	0,25	64000	240
Sumas	240		5.400		201.000	

Como vemos, la mediana está en el intervalo [10 – 30].

$$Me = L_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i} = 10 + 20 \frac{120 - 80}{90} = 18,89$$

La mediana indica que, una vez ordenados los individuos de la muestra de menor a mayor, aquel que ocupa la posición central consume 18,89 € de combustible semanalmente.

Ejercicio 15

a) Tabla tipo III:

Asistencia al cine	Nº personas
[0–10]	20
[10–20]	10
[20–35]	6
[35–50]	4

b) Estudio acerca de la medida de tendencia central más representativa.

Este estudio podríamos realizarlo analizando los siguientes aspectos:

- b1) Comprobar si la media aritmética es representativa, mediante el coeficiente de variación de Pearson.
- b2) Comprobar si la mediana es más o menos representativa que la media aritmética, comparando la desviación media con la desviación mediana.
- b3) Comprobar si la mediana puede ser la medida de tendencia central más representativa, calculando el coeficiente de asimetría de Pearson.
- b4) Comprobar si la moda es la medida de tendencia central más representativa, mediante el índice de Gini.

Todo ello aplicado a una población.

b1) Coeficiente de variación de Pearson:

$$V = \frac{\sigma}{\mu} * 100$$

Calcularemos los datos necesarios.

Media aritmética:

$$\mu = \frac{\sum x_i \cdot n_i}{N} = \frac{585}{40} = 14,625$$

Clases	n_i	x_i	$x_i n_i$
[0–10]	20	5	100
[10–20]	10	15	150
[20–35]	6	27,5	165
[35–50]	4	42,5	170
Sumas	40		585

Desviación típica:

Es la raíz cuadrada de la varianza. Calculamos primero la varianza:

$$\sigma = \frac{\sum X_i^2 \cdot n_i}{N} - \mu^2 = \frac{14512}{40} - 14,625^2 = 148,92$$

Clases	n_i	x_i	$x_i n_i$	$X_i^2 n_i$
[0–10]	20	5	100	500
[10–20]	10	15	150	2250
[20–35]	6	27,5	165	4537,5
[35–50]	4	42,5	170	7225
Sumas	40		585	14.512,5

$$\sigma = \sqrt{\sigma^2} = \sqrt{148,92} = 12,20$$

Por tanto, el coeficiente de variación será el siguiente:

$$V = \frac{\sigma}{\mu} * 100$$

$$V = \frac{12,20}{14,625} * 100 = 83,42\%$$

La media aritmética no es nada representativa, ya que es superior al 60%.

b2) ¿Sería más representativa la mediana?

Para responder a esta cuestión debemos calcular la desviación media y compararla con la desviación mediana. Si esta última fuera inferior, sería más representativa la mediana.

Desviación media:

$$DM = \frac{\sum |X_i - \mu| \cdot n_i}{N} = \frac{385}{40} = 9,625$$

Clases	n_i	x_i	$x_i n_i$	$ X_i - \bar{X} n_i$	N_i
[0-10]	20	5	100	192,5	20
[10-20]	10	15	150	3,75	30
[20-35]	6	27,5	165	77,25	36
[35-50]	4	42,5	170	111,5	40
Sumas	40		585	385	

Desviación mediana:

Calcularemos primero la mediana. El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($40/2 = 20$). Como vemos, la mediana está en el intervalo [10 - 20].

$$Me = L_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i} = 10 + 10 \frac{20 - 20}{10} = 10$$

La desviación mediana la calcularemos con la siguiente fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{385}{40} = 9,625$$

Clases	n_i	x_i	$x_i n_i$	$ X_i - \bar{X} n_i$	$ X_i - Me n_i$
[0-10]	20	5	100	192,5	100
[10-20]	10	15	150	3,75	50
[20-35]	6	27,5	165	77,25	105
[35-50]	4	42,5	170	111,5	130
Sumas	40		585	385	385

Como vemos, la desviación mediana es igual a la desviación media. Por ello, podemos decir que la media aritmética y la mediana serían igual de representativas de esta distribución.

b3) Asimetría:

Calcularemos el coeficiente de asimetría de Pearson, simbolizado por A, con la siguiente fórmula:

$$A = \frac{\mu - Mo}{\sigma}$$

Para ello, nos falta calcular la moda.

Moda:

Como se trata de intervalos irregulares hemos de calcular las alturas.

Clases	n_i	h_i
[0–10]	20	2
[10–20]	10	1
[20–35]	6	0,4
[35–50]	4	0,267
Sumas	40	

El intervalo que contiene a la moda será el de mayor altura (h_i). En nuestro caso, se trata del intervalo [0 – 10]. Como es el primer intervalo y además son intervalos de diferente amplitud, tenemos que aplicar la fórmula específica para este caso:

$$Mo = L_{i-1} + a_i \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) - (h_i - h_{i+1})}$$

$$Mo = 0 + 10 \frac{2 - 0}{(2 - 0) + (2 - 1)} = 6,67$$

Con ello, el coeficiente de asimetría de Pearson resulta:

$$A = \frac{\mu - Mo}{\sigma} = \frac{14,625 - 6,67}{12,20} = 0,65$$

Como es igual a 0,65 podemos decir que existe asimetría, pero no excesiva. Por ello, no podemos afirmar que la medida de tendencia central más significativa de esta distribución sea la mediana.

b4) Concentración:

Calcularemos el índice de Gini.

$$G = 1 - \frac{\sum Q_i}{\sum P_i} = 1 - \frac{230,76}{315} = 0,2674$$

Clases	n_i	x_i	$x_i n_i$	f_i	P_i	Q_i
[0–10]	20	5	100	0,5	50	17,09
[10–20]	10	15	150	0,25	75	42,73
[20–35]	6	27,5	165	0,15	90	70,94
[35–50]	4	42,5	170	0,10	100	100
Sumas	40		585	1	315	230,76

P_i = frecuencia relativa acumulada, multiplicada por 100

Q_i = cada producto $X_i n_i$, haciéndolo relativo y acumulado, multiplicado por 100.

Existe poca concentración (diríamos que hay mucha concentración cuando $G > 0,75$). Por tanto, la moda, según este estudio, no sería la medida de tendencia central más representativa de esta distribución.

Con todo lo analizado hasta aquí, podemos concluir que la moda no es el promedio más representativo de la distribución. La media aritmética y la mediana serían igualmente representativas, aunque tampoco serían ideales.

Ante un caso como este sería conveniente realizar el estudio de la muestra utilizando más intervalos, de manera que podamos obtener mayor información.

Podríamos comprobarlo utilizando 5 intervalos regulares.

En primer lugar, hemos de calcular el rango, que es la diferencia entre el mayor valor y el menor:

$$\text{Rango} = 50 - 0 = 50$$

Como queremos determinar 5 clases de la misma amplitud, dividiremos el rango entre el número de clases para calcular esa amplitud:

$$\text{Amplitud} = \text{Rango} / N^{\circ} \text{ clases} = 50 / 5 = 10$$

Clases	Frecuencia
[0 – 10]	20
[10 – 20]	10
[20 – 30]	5
[30 – 40]	3
[40 – 50]	2

Haremos con esta tabla el mismo estudio que el realizado anteriormente:

bI) Coeficiente de variación de Pearson:

$$V = \frac{\sigma}{\mu} * 100$$

Calcularemos los datos necesarios.

Media aritmética:

$$\mu = \frac{\sum x_i n_i}{N} = \frac{570}{40} = 14,25$$

Clases	n_i	x_i	$x_i n_i$
[0 – 10]	20	5	100
[10 – 20]	10	15	150
[20 – 30]	5	25	125
[30 – 40]	3	35	105
[40 – 50]	2	45	90
Sumas	40		570

Desviación típica:

Es la raíz cuadrada de la varianza. Calculamos primero la varianza:

$$\sigma^2 = \frac{\sum X_i^2 \cdot n_i}{N} - \mu^2 = \frac{13.600}{40} - 14,25^2 = 136,94$$

Clases	n_i	x_i	$x_i n_i$	$X^2_i n_i$
[0 – 10]	20	5	100	500
[10 – 20]	10	15	150	2250
[20 – 30]	5	25	125	3125
[30 – 40]	3	35	105	3675
[40 – 50]	2	45	90	4050
Sumas	40		570	13.600

$$\sigma = \sqrt{\sigma^2} = \sqrt{136,94} = 11,70$$

Por tanto, el coeficiente de variación será el siguiente:

$$V = \frac{\sigma}{\mu} * 100$$

$$V = \frac{11,70}{14,25} * 100 = 82,11\%$$

La media aritmética no es nada representativa, ya que es superior al 60%.

b2) ¿Sería más representativa la mediana?

Para responder a esta cuestión debemos calcular la desviación media y compararla con la desviación mediana. Si esta última fuera inferior, sería más representativa la mediana.

Desviación media:

$$DM = \frac{\sum |X_i - \mu| \cdot n_i}{N} = \frac{370}{40} = 9,25$$

Clases	n_i	x_i	$x_i n_i$	$ X_i - \bar{X} n_i$	N_i
[0 – 10]	20	5	100	185	20
[10 – 20]	10	15	150	7,5	30
[20 – 30]	5	25	125	53,75	35
[30 – 40]	3	35	105	62,25	38
[40 – 50]	2	45	90	61,5	40
Sumas	40		570	370	

Desviación mediana:

Calcularemos primero la mediana. El intervalo que contiene a la mediana será el primero cuya frecuencia absoluta acumulada supere a $N/2$ ($40/2 = 20$). Como vemos, la mediana está en el intervalo $[10 - 20]$.

$$Me = L_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i} = 10 + 10 \frac{20 - 20}{10} = 10$$

La desviación mediana la calcularemos con la siguiente fórmula:

$$DMe = \frac{\sum |X_i - Me| \cdot n_i}{N} = \frac{370}{40} = 9,25$$

Clases	n_i	x_i	$x_i n_i$	$ X_i - \bar{X} n_i$	$ X_i - Me n_i$
$[0 - 10]$	20	5	100	185	100
$[10 - 20]$	10	15	150	7,5	50
$[20 - 30]$	5	25	125	53,75	75
$[30 - 40]$	3	35	105	62,25	75
$[40 - 50]$	2	45	90	61,5	70
Sumas	40		570	370	370

Como vemos, la desviación mediana es igual a la desviación media. Por ello, podemos decir que la media aritmética y la mediana serían igual de representativas de esta distribución.

b3) Asimetría:

Calcularemos el coeficiente de asimetría de Pearson, simbolizado por A , con la siguiente fórmula:

$$A = \frac{\mu - Mo}{\sigma}$$

Para ello, nos falta calcular la moda.

Moda:

Como se trata de intervalos irregulares hemos de calcular las alturas.

Clases	n_i	h_i
[0 – 10]	20	2
[10 – 20]	10	1
[20 – 30]	5	0,5
[30 – 40]	3	0,3
[40 – 50]	2	0,2
Sumas	40	

El intervalo que contiene a la moda será el de mayor altura (h_i). En nuestro caso, se trata del intervalo [0 – 10]. Como es el primer intervalo y además son intervalos de diferente amplitud, tenemos que aplicar la fórmula específica para este caso:

$$Mo = L_{i-1} + a_i \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})}$$

$$Mo = 0 + 10 \frac{2 - 0}{(2 - 0) + (2 - 1)} = 6,67$$

Con ello, el coeficiente de asimetría de Pearson resulta:

$$A = \frac{\mu - Mo}{\sigma} = \frac{14,25 - 6,67}{11,70} = 0,648$$

Como es ligeramente inferior a 0,65 podemos decir que existe poca asimetría. Por ello, podemos afirmar que la medida de tendencia central más significativa de esta distribución podría ser la mediana.

b4) Concentración:

Calcularemos el índice de Gini.

$$G = 1 - \frac{\sum Q_i}{\sum P_i} = 1 - \frac{311,4}{407,5} = 0,2358$$

Clases	n_i	x_i	$x_i n_i$	f_i	P_i	Q_i
[0 – 10]	20	5	100	0,5	50	17,54
[10 – 20]	10	15	150	0,25	75	43,86
[20 – 30]	5	25	125	0,125	87,5	65,79
[30 – 40]	3	35	105	0,075	95	84,21
[40 – 50]	2	45	90	0,05	100	100
Sumas	40		570	1	407,5	311,4

P_i = frecuencia relativa acumulada, multiplicada por 100

Q_i = cada producto $X_i n_i$, haciéndolo relativo y acumulado, multiplicado por 100.

Existe poca concentración (diríamos que hay mucha concentración cuando $G > 0,75$). Por tanto, la moda, según este estudio, no sería la medida de tendencia central más representativa de esta distribución.

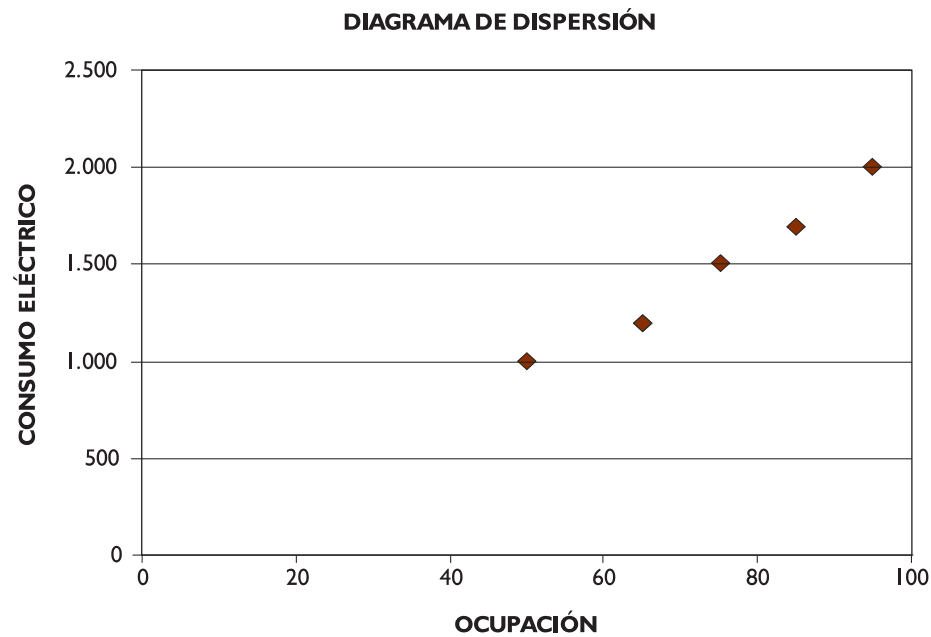
Según el estudio, parece que sea la mediana la medida más conveniente de tendencia central en esta distribución.

Soluciones ejercicios. Tema 8

Ejercicio I

a) Diagrama de dispersión

% OCUP.	x_i	y_j	n_{ij}
40 – 60	50	1.000	5
60 – 70	65	1.200	15
70 – 80	75	1.500	25
80 – 90	85	1.700	10
90 – 100	95	2.000	5



b) Recta de regresión

Calcularemos los valores “a” y “b” de la recta:

$$a = \mu_Y - b\mu_X$$

$$b = \frac{\sigma_{XY}}{\sigma_X^2}$$

Media de X:

$$\mu_X = \frac{\sum x_i n_i}{N} = \frac{4425}{60} = 73,75$$

Clases	n_i	x_i	$x_i n_i$
40 – 60	5	50	250
60 – 70	15	65	975
70 – 80	25	75	1875
80 – 90	10	85	850
90 – 100	5	95	475
Sumas	60		4.425

Media de Y:

$$\mu_Y = \frac{\sum Y_j n_j}{N} = \frac{87500}{60} = 1.458,33$$

Y_j	n_j	$Y_j n_j$
1.000	5	5000
1.200	15	18000
1.500	25	37500
1.700	10	17000
2.000	5	10000
Sumas	60	87.500

Varianza de X:

$$\sigma_x^2 = \sum \frac{X_i^2 \cdot n_i}{N} - \mu_x^2 = \frac{333875}{60} - 73,75^2 = 125,52$$

Clases	n_i	x_i	$x_i n_i$	$X_i^2 n_i$
40 – 60	5	50	250	12500
60 – 70	15	65	975	63375
70 – 80	25	75	1875	140625
80 – 90	10	85	850	72250
90 – 100	5	95	475	45125
Sumas	60		4.425	333.875

Covarianza:

$$\sigma_{xy} = \frac{\sum \sum X_i Y_j n_{ij}}{N} - \mu_x \cdot \mu_y = \frac{6627500}{60} - 73,75 \cdot 1458,33 = 2.906,5$$

x_i	y_j	n_{ij}	$x_i \cdot y_j \cdot n_{ij}$
50	1.000	5	250000
65	1.200	15	1170000
75	1.500	25	2812500
85	1.700	10	1445000
95	2.000	5	950000
Sumas		60	6.627.500

Con ello, tenemos lo siguiente:

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{2906,5}{125,52} = 23,16$$

$$a = \mu_y - b\mu_x = 1458,33 - 23,16 \cdot 73,75 = -249,72$$

La recta de regresión será:

$$Y = -249,72 + 23,16 X$$

c) Coeficiente de determinación:

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2}$$

Hemos de calcular la varianza residual y la varianza de Y

Varianza residual:

Calculamos primero los valores de Y_i^* , que son lo que obtendremos de la ecuación de la recta cuando sustituimos los valores observados de X_i .

x_i	y_j	n_{ij}	Y_i^*	$(Y_j - Y_i^*)^2 \cdot n_{ij}$
50	1.000	5	908,28	42062,79
65	1.200	15	1255,68	46503,94
75	1.500	25	1487,28	4044,96
85	1.700	10	1718,88	3564,54
95	2.000	5	1950,48	12261,15
Sumas		60		108.437,38

$$S^2_e = \frac{\sum \sum (Y_j - Y_i^*)^2 \cdot n_{ij}}{N} = \frac{108.437,38}{60} = 1.807,29$$

Varianza de Y:

$$\sigma_Y^2 = \sum \frac{Y_j^2 \cdot n_j}{N} - \mu_Y^2 = \frac{131750000}{60} - 1458,33^2 = 69,106,94$$

y_j	n_j	$y_j n_j$	$y_j^2 n_j$
1.000	5	5000	5000000
1.200	15	18000	21600000
1.500	25	37500	56250000
1.700	10	17000	28900000
2.000	5	10000	20000000
Sumas	60	87.500	131.750.000

Con estos resultados, el coeficiente de determinación será el siguiente:

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2} = 1 - \frac{1807,29}{69106,94} = 0,9738$$

Podemos decir que la recta es una función que explica adecuadamente la relación entre la variable independiente X (ocupación) y la variable dependiente Y (consumo eléctrico), puesto que R^2 es superior a 0,90.

d) Coeficiente de correlación lineal

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Como las desviaciones típicas son la raíz cuadrada de las varianzas y éstas ya están calculadas para las dos variables, tenemos lo siguiente:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{125,52} = 11,20$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{69106,94} = 262,88$$

Por tanto, el coeficiente de correlación lineal será:

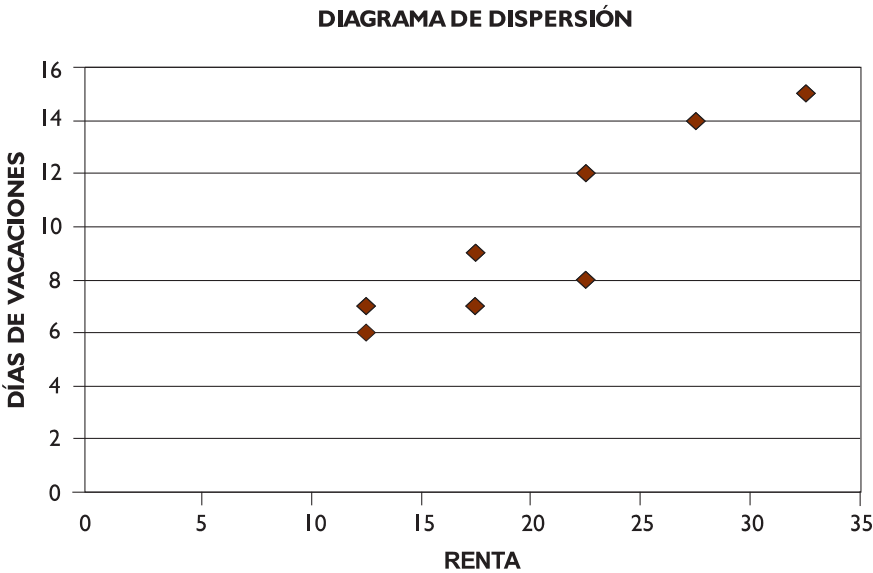
$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{2906,5}{11,20 \cdot 262,88} = 0,987$$

Al ser superior a 0,85 podemos decir que existe relación lineal entre las variables y esta relación es directa (cuando una crece, la otra también).

Ejercicio 2

a) Diagrama de dispersión

Renta	x_i	y_j	n_{ij}
10 – 15	12,5	6	50
10 – 15	12,5	7	120
15 – 20	17,5	7	180
15 – 20	17,5	9	220
20 – 25	22,5	8	150
20 – 25	22,5	12	140
25 – 30	27,5	14	100
30 – 35	32,5	15	40



b) Recta de regresión

Calcularemos los valores “a” y “b” de la recta:

$$a = \mu_y - b\mu_x$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

Media de X:

$$\mu_x = \frac{\sum x_j n_j}{N} = \frac{19700}{1000} = 19,70$$

Renta	x _i	n _i	x _i n _i
10 – 15	12,5	50	625
10 – 15	12,5	120	1500
15 – 20	17,5	180	3150
15 – 20	17,5	220	3850
20 – 25	22,5	150	3375
20 – 25	22,5	140	3150
25 – 30	27,5	100	2750
30 – 35	32,5	40	1300
Sumas		1.000	19.700

Media de Y:

$$\mu_y = \frac{\sum Y_j n_j}{N} = \frac{9260}{1000} = 9,26$$

y _i	n _i	y _i n _i
6	50	300
7	120	840
7	180	1260
9	220	1980
8	150	1200
12	140	1680
14	100	1400
15	40	600
Sumas	1.000	9.260

Varianza de X:

$$\sigma_X^2 = \frac{\sum X_i^2 \cdot n_i}{N} - \mu_X^2 = \frac{413750}{1000} - 19,70^2 = 25,66$$

Renta	x_i	n_i	$x_i n_i$	$X_i^2 n_i$
10 – 15	12,5	50	625	7812,5
10 – 15	12,5	120	1500	18750
15 – 20	17,5	180	3150	55125
15 – 20	17,5	220	3850	67375
20 – 25	22,5	150	3375	75937,5
20 – 25	22,5	140	3150	70875
25 – 30	27,5	100	2750	75625
30 – 35	32,5	40	1300	42250
Sumas		1.000	19.700	413.750

Covarianza:

$$\sigma_{XY} = \frac{\sum \sum X_i Y_j n_{ij}}{N} - \mu_X \cdot \mu_Y = \frac{193750}{1000} - 19,7 \cdot 9,26 = 11,328$$

x_i	y_j	n_{ij}	$x_i \cdot y_j \cdot n_{ij}$
12,5	6	50	3750
12,5	7	120	10500
17,5	7	180	22050
17,5	9	220	34650
22,5	8	150	27000
22,5	12	140	37800
27,5	14	100	38500
32,5	15	40	19500
Sumas		1.000	193.750

Con ello, tenemos lo siguiente:

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{11,328}{25,66} = 0,441465$$

La recta de regresión será:

$$Y = 0,5631 + 0,441465 \cdot X$$

c) Coeficiente de determinación

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2}$$

Hemos de calcular la varianza residual y la varianza de Y.

Varianza residual

Calculamos primero los valores de Y_i^* , que son lo que obtendremos de la ecuación de la recta cuando sustituimos los valores observados de X_i .

x_i	y_j	n_{ij}	Y_i^*	$(Y_j - Y_i^*)^2 \cdot n_{ij}$
12,5	6	50	6,08	0,32
12,5	7	120	6,08	101,568
17,5	7	180	8,29	299,538
17,5	9	220	8,29	110,902
22,5	8	150	10,50	937,5
22,5	12	140	10,50	315
27,5	14	100	12,70	169
32,5	15	40	14,91	0,324
Sumas		1.000		1.934,152

$$S^2_e = \frac{\sum \sum (Y_j - Y_i^*)^2 \cdot n_{ij}}{N} = \frac{1934,152}{1000} = 1,934$$

Varianza de Y:

$$\sigma_Y^2 = \frac{\sum Y_i^2 \cdot n_i}{N} - \mu_Y^2 = \frac{92680}{1000} - 9,26^2 = 6,9324$$

y_j	n_j	$y_j n_j$	$y_j^2 n_j$
6	50	300	1800
7	120	840	5880
7	180	1260	8820
9	220	1980	17820
8	150	1200	9600
12	140	1680	20160
14	100	1400	19600
15	40	600	9000
Sumas	1.000	9.260	92.680

Con estos resultados, el coeficiente de determinación será el siguiente:

$$R_2 = 1 - \frac{S^2_e}{\sigma_Y^2} = 1 - \frac{1,934}{6,9324} = 0,7210$$

Podemos decir que la relación entre la variable independiente X (RENTA) y la variable dependiente Y (DÍAS DE VACACIONES) no sigue una relación lineal, puesto que R^2 es inferior a 0,75. Quizás pueda existir relación, pero no con una recta.

d) Coeficiente de correlación lineal

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Como las desviaciones típicas son la raíz cuadrada de las varianzas y éstas ya están calculadas para las dos variables, tenemos lo siguiente:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{25,66} = 5,066$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{69106,94} = 262,88$$

Por tanto, el coeficiente de correlación lineal será:

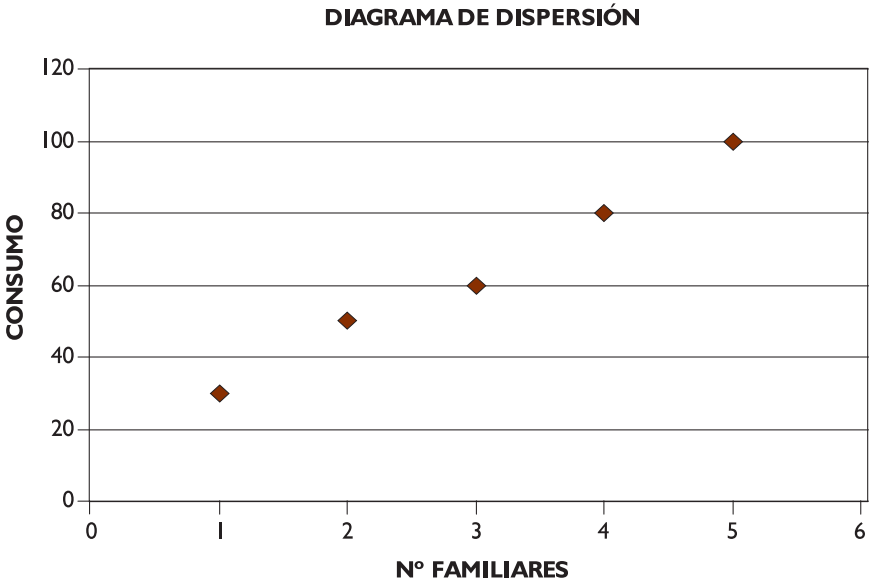
$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{11,328}{5,066 \cdot 2,633} = 0,8493$$

Al ser inferior a 0,85 podemos decir que no existe relación lineal entre las variables.

Ejercicio 3

a) Diagrama de dispersión

x_i	y_j	n_{ij}
1	30	100
2	50	200
3	60	300
4	80	150
5	100	50



b) Recta de regresión

Calcularemos los valores “a” y “b” de la recta:

$$a = \mu_Y - b\mu_X$$

$$b = \frac{\sigma_{XY}}{\sigma_X^2}$$

Media de X:

$$\mu_x = \frac{\sum x_i n_i}{N} = \frac{2250}{800} = 2,8125$$

x_i	n_i	$x_i n_i$
1	100	100
2	200	400
3	300	900
4	150	600
5	50	250
Sumas	800	2.250

Media de Y:

$$\mu_y = \frac{\sum Y_j n_j}{N} = \frac{48000}{800} = 60$$

y_i	n_i	$y_i n_i$
30	100	3000
50	200	10000
60	300	18000
80	150	12000
100	50	5000
Sumas	800	48.000

Varianza de X:

$$\sigma_x^2 = \frac{\sum x_i^2 \cdot n_i}{N} - \mu_x^2 = \frac{7250}{800} - 2,8125^2 = 1,15234$$

x_i	n_i	$x_i n_i$	$X^2_i n_i$
1	100	100	100
2	200	400	800
3	300	900	2700
4	150	600	2400
5	50	250	1250
Sumas	800	2.250	7.250

Covarianza:

$$\sigma_{XY} = \frac{\sum \sum X_i Y_j n_{ij}}{N} - \mu_X \cdot \mu_Y = \frac{150000}{800} - 2,8125 \cdot 60 = 18,75$$

x_i	y_j	n_{ij}	$x_i \cdot y_j \cdot n_{ij}$
1	30	100	3000
2	50	200	20000
3	60	300	54000
4	80	150	48000
5	100	50	25000
Sumas		800	150.000

Con ello, tenemos lo siguiente:

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{18,75}{1,15234} = 16,2712$$

$$a = \mu_Y - b\mu_X = 60 - 16,2712 \cdot 2,8125 = 14,2373$$

La recta de regresión será:

$$Y = 14,2373 + 16,2712 \cdot X$$

c) Coeficiente de determinación

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2}$$

Hemos de calcular la varianza residual y la varianza de Y .

Varianza residual:

Calculamos primero los valores de Y_i^* , que son lo que obtendremos de la ecuación de la recta cuando sustituimos los valores observados de X_i .

x_i	y_j	n_{ij}	Y_i^*	$(Y_j - Y_i^*)^2 \cdot n_{ij}$
1	30	100	30,51	25,85
2	50	200	46,78	2074,12
3	60	300	63,05	2792,30
4	80	150	79,32	68,95
5	100	50	95,59	970,99
Sumas		800		5.932,2

$$S^2_e = \frac{\sum \sum (Y_j - Y_i^*)^2 \cdot n_{ij}}{N} = \frac{5932,2}{800} = 7,41525$$

Varianza de Y :

$$\sigma_Y^2 = \frac{\sum Y_j^2 \cdot n_j}{N} - \mu_Y^2 = \frac{3130000}{800} - 60^2 = 312,5$$

y_j	n_j	$y_j n_j$	$y_j^2 n_j$
30	100	3000	90000
50	200	10000	500000
60	300	18000	1080000
80	150	12000	960000
100	50	5000	500000
Sumas	800	48.000	3.130.000

Con estos resultados, el coeficiente de determinación será el siguiente:

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2} = 1 - \frac{7,41525}{312,5} = 0,9763$$

Podemos decir que la relación entre la variable independiente X (Nº DE FAMILIARES) y la variable dependiente Y (CONSUMO) sigue una relación lineal, puesto que R^2 es superior a 0,90.

d) Coeficiente de correlación lineal

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Como las desviaciones típicas son la raíz cuadrada de las varianzas y éstas ya están calculadas para las dos variables, tenemos lo siguiente:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{1,15234} = 1,07347$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{312,5} = 17,6777$$

Por tanto, el coeficiente de correlación lineal será:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{18,75}{1,07347 \cdot 17,6777} = 0,988$$

Al ser superior a 0,85 podemos decir que existe relación lineal entre las variables y que esta es directa.

- e) El consumo de una familia de 6 miembros lo podríamos obtener sustituyendo $x=6$ en la recta de regresión.

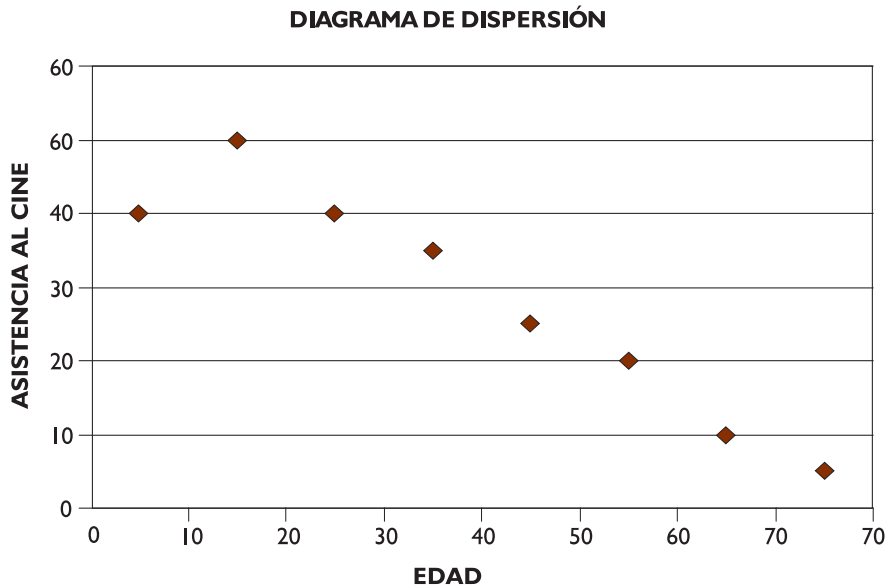
$$Y = 14,2373 + 16,2712 \cdot X$$

$$Y = 14,2373 + 16,2712 \cdot 6 = 111,86 \text{ €}$$

Ejercicio 4

a) Diagrama de dispersión

Edad	x_i	y_j	n_{ij}
0 – 10	5	40	50
10 – 20	15	50	100
20 – 30	25	40	150
30 – 40	35	35	200
40 – 50	45	25	200
50 – 60	55	20	100
60 – 70	65	10	100
70 – 80	75	5	100



b) Recta de regresión

Calcularemos los valores “a” y “b” de la recta:

$$a = \mu_Y - b\mu_X$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

Media de X:

$$\mu_x = \frac{\sum x_i n_i}{N} = \frac{41000}{1000} = 41$$

Edad	x_i	n_i	$x_i n_i$
0 – 10	5	50	250
10 – 20	15	100	1500
20 – 30	25	150	3750
30 – 40	35	200	7000
40 – 50	45	200	9000
50 – 60	55	100	5500
60 – 70	65	100	6500
70 – 80	75	100	7500
Sumas		1.000	41.000

Media de Y:

$$\mu_y = \frac{\sum Y_j n_j}{N} = \frac{28500}{1000} = 28,5$$

y_i	n_j	$y_i n_j$
40	50	2000
50	100	5000
40	150	6000
35	200	7000
25	200	5000
20	100	2000
10	100	1000
5	100	500
Sumas	1.000	28.500

Varianza de X:

$$\sigma_X^2 = \frac{\sum X_i^2 \cdot n_i}{N} - \mu_X^2 = \frac{2055000}{1000} - 41^2 = 374$$

Edad	x_i	n_i	$x_i n_i$	$X_i^2 n_i$
0 – 10	5	50	250	1250
10 – 20	15	100	1500	22500
20 – 30	25	150	3750	93750
30 – 40	35	200	7000	245000
40 – 50	45	200	9000	405000
50 – 60	55	100	5500	302500
60 – 70	65	100	6500	422500
70 – 80	75	100	7500	562500
Sumas		1.000	41.000	2.055.000

Covarianza:

$$\sigma_{XY} = \frac{\sum \sum X_i Y_j n_{ij}}{N} - \mu_X \cdot \mu_Y = \frac{917500}{1000} - 41 \cdot 28,5 = -251$$

x_i	y_j	n_{ij}	$x_i \cdot y_j \cdot n_{ij}$
5	40	50	10000
15	50	100	75000
25	40	150	150000
35	35	200	245000
45	25	200	225000
55	20	100	110000
65	10	100	65000
75	5	100	37500
Sumas		1.000	917.500

Con ello, tenemos lo siguiente:

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{-251}{374} = -0,671123$$

$$a = \mu_Y - b\mu_X = 28,05 - (-0,671123) \cdot 41 = 56,016043$$

La recta de regresión será:

$$Y = 56,016043 - 0,671123 \cdot X$$

c) Coeficiente de determinación

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2}$$

Hemos de calcular la varianza residual y la varianza de Y

Varianza residual:

Calculamos primero los valores de Y^*_i , que son lo que obtendremos de la ecuación de la recta cuando sustituimos los valores observados de X_i .

x_i	y_j	n_{ij}	Y^*_i	$(Y_j - Y^*_i)^2 \cdot n_{ij}$
5	40	50	52,66	8014,32
15	50	100	45,95	1640,90
25	40	150	39,24	87,10
35	35	200	32,53	1223,41
45	25	200	25,82	133,01
55	20	100	19,10	80,23
65	10	100	12,39	572,67
75	5	100	5,68	46,49
Sumas		1.000		11.798,1

$$S^2_e = \frac{\sum \sum (Y_j - Y^*_i)^2 \cdot n_{ij}}{N} = \frac{11798,1}{1000} = 11,7981$$

Varianza de Y:

$$\sigma_Y^2 = \frac{\sum Y_i^2 \cdot n_{ij}}{N} - \mu_Y^2 = \frac{992500}{1000} - 28,5^2 = 180,25$$

y_j	n_j	$y_j n_j$	$y_j^2 n_j$
40	50	2000	80000
50	100	5000	250000
40	150	6000	240000
35	200	7000	245000
25	200	5000	125000
20	100	2000	40000
10	100	1000	10000
5	100	500	2500
Sumas	1.000	28.500	992.500

Con estos resultados, el coeficiente de determinación será el siguiente:

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2} = 1 - \frac{11,7981}{180,25} = 0,9345$$

Podemos decir que la relación entre la variable independiente X (EDAD) y la variable dependiente Y (ASISTENCIA AL CINE) sigue claramente una relación lineal, puesto que R^2 es superior a 0,90.

d) Coeficiente de correlación lineal

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Como las desviaciones típicas son la raíz cuadrada de las varianzas y éstas ya están calculadas para las dos variables, tenemos lo siguiente:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{374} = 19,3391$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{180,25} = 13,4257$$

Por tanto, el coeficiente de correlación lineal será:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{-251}{19,3391 \cdot 13,4257} = -0,9667$$

Al ser inferior a $-0,85$ podemos decir que existe relación lineal entre las variables y que es una relación inversa (cuando aumenta X, disminuye Y).

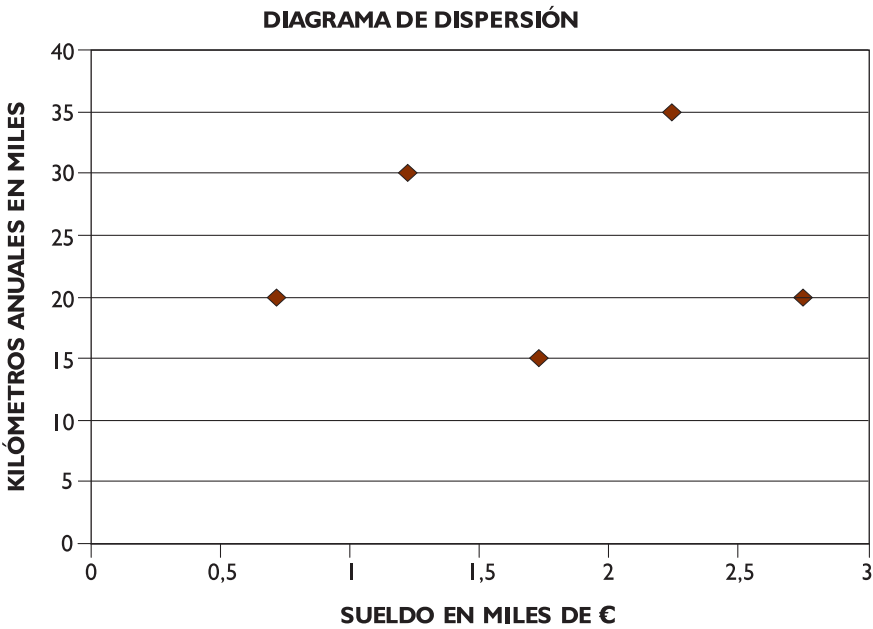
Ejercicio 5

a) Diagrama de dispersión

La tabla quedaría de la siguiente forma:

Sueldo	x_i	y_j	n_{ij}
500 – 1000	750	20.000	150
1000 – 1500	1250	30.000	200
1500 – 2000	1750	15.000	250
2000 – 2500	2250	35.000	100
2500 – 3000	2750	20.000	200

Para facilitar los cálculos expresaremos el sueldo en miles de euros y los kilómetros anuales en miles.



La tabla quedaría como la siguiente:

Sueldo	x_i	y_j	n_{ij}
0,50 – 1	0,750	20	150
1 – 1,5	1,25	30	200
1,5 – 2	1,75	15	250
2 – 2,5	2,25	35	100
2,5 – 3	2,75	20	200

b) Recta de regresión

Calcularemos los valores “a” y “b” de la recta:

$$a = \mu_Y - b\mu_X$$

$$b = \frac{\sigma_{XY}}{\sigma_X^2}$$

Media de X:

$$\mu_X = \frac{\sum x_i n_i}{N} = \frac{1575}{900} = 1,75$$

Sueldo	x_i	n_i	$x_i n_i$
0 – 10	0,750	150	112,5
10 – 20	1,25	200	250
20 – 30	1,75	250	437,5
30 – 40	2,25	100	225
40 – 50	2,75	200	550
Sumas		900	1.575

Media de Y:

$$\mu_Y = \frac{\sum Y_j n_j}{N} = \frac{20250}{900} = 22,5$$

y_i	n_j	$y_i n_j$
20	150	3000
30	200	6000
15	250	3750
35	100	3500
20	200	4000
Sumas	900	20.250

Varianza de X:

$$\sigma_x^2 = \frac{\sum X_i^2 \cdot n_i}{N} - \mu_x^2 = \frac{3181,25}{900} - 1,75^2 = 0,472222$$

Sueldo	x_i	n_i	$x_i n_i$	$X_i^2 n_i$
0 – 10	0,750	150	112,5	84,375
10 – 20	1,25	200	250	312,5
20 – 30	1,75	250	437,5	765,625
30 – 40	2,25	100	225	506,25
40 – 50	2,75	200	550	1512,5
Sumas		900	1.575	3.181,25

Covarianza:

$$\sigma_{xy} = \frac{\sum \sum X_i Y_j n_{ij}}{N} - \mu_x \cdot \mu_y = \frac{35,187,5}{900} - 1,75 \cdot 22,5 = -0,277778$$

x_i	y_j	n_{ij}	$x_i \cdot y_j \cdot n_{ij}$
0,750	20	150	2250
1,25	30	200	7500
1,75	15	250	6562,5
2,25	35	100	7875
2,75	20	200	11000
Sumas		900	35.187,5

Con ello, tenemos lo siguiente:

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{-0,277778}{0,472222} = -0,588236$$

$$a = \mu_Y - b\mu_X = 22,5 - (-0,588236) \cdot 1,75 = 23,529413$$

La recta de regresión será:

$$Y = 23,529413 - 0,588236 \cdot X$$

c) Coeficiente de determinación

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2}$$

Hemos de calcular la varianza residual y la varianza de Y.

Varianza residual:

Calculamos primero los valores de Y_i^* , que son lo que obtendremos de la ecuación de la recta cuando sustituimos los valores observados de X_i .

x_i	y_i	n_{ij}	Y_i^*	$(Y_i - Y_i^*)^2 \cdot n_{ij}$
0,750	20	150	23,09	1430,58
1,25	30	200	22,79	10384,95
1,75	15	250	22,50	14062,50
2,25	35	100	22,21	16368,94
2,75	20	200	21,91	730,97
Sumas		900		42977,94

$$S^2_e = \frac{\sum \sum (Y_i - Y_i^*)^2 \cdot n_{ij}}{N} = \frac{42977,94}{900} = 47,753266$$

Varianza de Y:

$$\sigma_Y^2 = \frac{\sum Y_i^2 \cdot n_{ij}}{N} - \mu_Y^2 = \frac{498750}{900} - 22,5^2 = 47,916667$$

y_j	n_j	$y_j n_j$	$y_j^2 n_j$
20	150	3000	60000
30	200	6000	180000
15	250	3750	56250
35	100	3500	122500
20	200	4000	80000
Sumas	900	20.250	498.750

Con estos resultados, el coeficiente de determinación será el siguiente:

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2} = 1 - \frac{47,753266}{47,916667} = 0,0034102$$

Podemos decir que la relación entre la variable independiente X (SUELDO) y la variable dependiente Y (KILÓMETROS ANUALES) no es una relación lineal, y posiblemente no exista entre ellas ningún tipo de relación, puesto que R^2 es muy cercano a 0.

d) Coeficiente de correlación lineal

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Como las desviaciones típicas son la raíz cuadrada de las varianzas y éstas ya están calculadas para las dos variables, tenemos lo siguiente:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{0,472222} = 0,687184$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{47,916667} = 6,922187$$

Por tanto, el coeficiente de correlación lineal será:

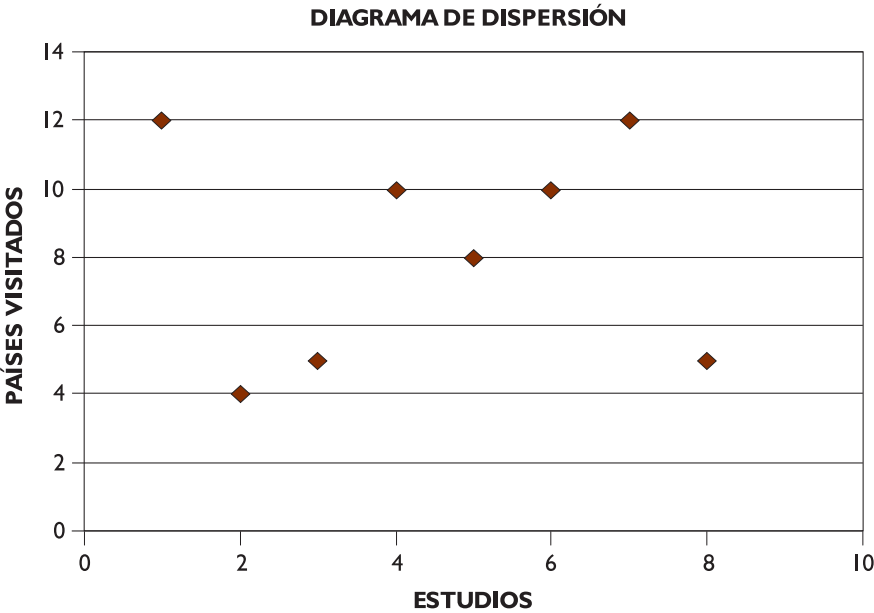
$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{-0,277778}{0,687184 \cdot 6,922187} = -0,0583958$$

Al ser tan cercano a 0 podemos decir que no existe relación lineal entre las variables.

Ejercicio 6

a) Diagrama de dispersión

x_i	y_j	n_{ij}
1	12	150
2	4	120
3	5	180
4	10	200
5	8	150
6	10	150
7	12	100
8	5	50



Podemos apreciar que la relación entre las variables, si existe, no va a ser mediante una recta.

b) Recta de regresión

Calcularemos los valores “a” y “b” de la recta:

$$a = \mu_Y - b\mu_X$$

$$b = \frac{\sigma_{XY}}{\sigma_X^2}$$

Media de X:

$$\mu_X = \frac{\sum x_i n_i}{N} = \frac{4480}{1100} = 4,0727273$$

x_i	n_i	$x_i n_i$
1	150	150
2	120	240
3	180	540
4	200	800
5	150	750
6	150	900
7	100	700
8	50	400
Sumas	1.100	4.480

Media de Y:

$$\mu_Y = \frac{\sum Y_i n_i}{N} = \frac{9330}{1100} = 8,481818$$

y_i	n_i	$y_i n_i$
12	150	1800
4	120	480
5	180	900
10	200	2000
8	150	1200
10	150	1500
12	100	1200
5	50	250
Sumas	1.100	9.330

Varianza de X:

$$\sigma_x^2 = \frac{\sum X_i^2 \cdot n_i}{N} - \mu_x^2 = \frac{22700}{1100} - 4,0727273^2 = 4,049256$$

x_i	n_i	$x_i n_i$	$X_i^2 n_i$
1	150	150	150
2	120	240	480
3	180	540	1620
4	200	800	3200
5	150	750	3750
6	150	900	5400
7	100	700	4900
8	50	400	3200
Sumas	1.100	4.480	22.700

Covarianza:

$$\sigma_{xy} = \frac{\sum \sum X_i Y_j n_{ij}}{N} - \mu_x \cdot \mu_y = \frac{38860}{1100} - 4,0727273 \cdot 8,481818 = 0,783141$$

x_i	y_j	n_{ij}	$x_i \cdot y_j \cdot n_{ij}$
1	12	150	1800
2	4	120	960
3	5	180	2700
4	10	200	8000
5	8	150	6000
6	10	150	9000
7	12	100	8400
8	5	50	2000
Sumas		1.100	38.860

Con ello, tenemos lo siguiente:

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{0,783141}{4,049256} = 0,1934036$$

$$a = \mu_Y - b\mu_X = 8,481818 - 0,1934036 \cdot 4,0727273 = 7,6941383$$

La recta de regresión será:

$$Y = 7,6941383 + 0,1934036 \cdot X$$

c) Coeficiente de determinación

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2}$$

Hemos de calcular la varianza residual y la varianza de Y.

Varianza residual:

Calculamos primero los valores de Y^*_i , que son lo que obtendremos de la ecuación de la recta cuando sustituimos los valores observados de X_i .

x_i	y_i	n_{ij}	Y^*_i	$(Y_j - Y^*_i)^2 \cdot n_{ij}$
1	12	150	7,89	2536,85
2	4	120	8,08	1998,49
3	5	180	8,27	1929,84
4	10	200	8,47	469,56
5	8	150	8,66	65,57
6	10	150	8,85	196,81
7	12	100	9,05	871,45
8	5	50	9,24	899,46
Sumas		1.100		8.968,028

$$S^2_e = \frac{\sum \sum (Y_j - Y^*_i)^2 \cdot n_{ij}}{N} = \frac{8968,028}{1100} = 8,152753$$

Varianza de Y:

$$\sigma_Y^2 = \frac{\sum Y_i^2 \cdot n_i}{N} - \mu_Y^2 = \frac{88270}{1100} - 8,481818^2 = 8,304215$$

y_j	n_j	$y_i n_i$	$y_j^2 n_j$
12	150	1800	21600
4	120	480	1920
5	180	900	4500
10	200	2000	20000
8	150	1200	9600
10	150	1500	15000
12	100	1200	14400
5	50	250	1250
Sumas	1.100	9.330	88.270

Con estos resultados, el coeficiente de determinación será el siguiente:

$$R^2 = 1 - \frac{S^2_e}{\sigma_Y^2} = 1 - \frac{8,152753}{8,304215} = 0,0182392$$

Podemos decir que la relación entre la variable independiente X (ESTUDIOS) y la variable dependiente Y (PAÍSES VISITADOS) no sigue una relación lineal, y posiblemente no exista entre ellas ningún tipo de relación, puesto que R^2 es muy cercano a 0.

d) Coeficiente de correlación lineal

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Como las desviaciones típicas son la raíz cuadrada de las varianzas y éstas ya están calculadas para las dos variables, tenemos lo siguiente:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{4,049256} = 2,012276$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{8,304215} = 2,881703$$

Por tanto, el coeficiente de correlación lineal será:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{0,783141}{2,012276 \cdot 2,881703} = 0.135053$$

Al ser cercano a 0 podemos decir que no existe relación lineal entre las variables.

Soluciones ejercicios. Tema 9

Ejercicio I

- a) Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema multiplicativo.

En primer lugar calculamos las medias móviles. Al ser un número impar de subperíodos, esas medias ya estarán centradas. Las incluimos en la tabla siguiente:

Años	Cuatrimestres		
	1º	2º	3º
2007		51,666667	55,00
2008	56,666667	58,333333	60,00
2009	60,00	61,666667	63,333333
2010	66,666667	68,333333	71,666667
2011	73,333333	75,00	

Como seguimos el método multiplicativo, dividiremos cada una de las observaciones de la tabla original entre estas medias móviles, obteniéndose el siguiente resultado:

Años	Cuatrimestres		
	1º	2º	3º
2007		1,161290	1,000000
2008	0,882353	1,114286	1,000000
2009	0,916667	1,054054	1,026316
2010	0,900000	1,097561	0,976744
2011	0,954545	1,066667	

A continuación calculamos las medias para cada uno de los trimestres:

Cuatrimestres		
1º	2º	3º
0,913391	1,098772	1,000765

La suma de esas medias debería dar como resultado 3, el número de subperíodos. Pero comprobamos que el resultado es 3,012928. Por ello, dividiremos cada una de las medias halladas entre su suma dividida por 3 ($3,012928 / 3 = 1,004309269$). El resultado es el siguiente:

Cuatrimestres		
1º	2º	3º
0,909472	1,094057	0,996471

Estos serán los índices normalizados.

b) Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema aditivo.

En primer lugar calculamos las medias móviles. Al ser un número impar de subperíodos, esas medias ya estarán centradas. Las incluimos en la tabla siguiente:

Años	Cuatrimestres		
	1º	2º	3º
2007		51,666667	55,00
2008	56,666667	58,333333	60,00
2009	60,00	61,666667	63,333333
2010	66,666667	68,333333	71,666667
2011	73,333333	75,00	

Como seguimos el método aditivo, restaremos a cada una de las observaciones de la tabla original estas medias móviles, obteniéndose el siguiente resultado:

	Cuatrimestres		
Años	1º	2º	3º
2007		8,333333	0,00
2008	-6,666667	6,666667	0,00
2009	-5,00	3,333333	1,666667
2010	-6,666667	6,666667	-1,666667
2011	-3,333333	5,00	

A continuación calculamos las medias para cada uno de los trimestres:

Cuatrimestres		
1º	2º	3º
-5,416667	6,00	0,00

La suma de esas medias debería dar como resultado 0. Pero comprobamos que el resultado es 0,583333. Por ello, restaremos a cada una de las medias halladas su suma dividida por 3 ($0,583333 / 3 = 0,194444$). El resultado es el siguiente:

Cuatrimestres		
1º	2º	3º
-5,611111	5,805556	-0,194444

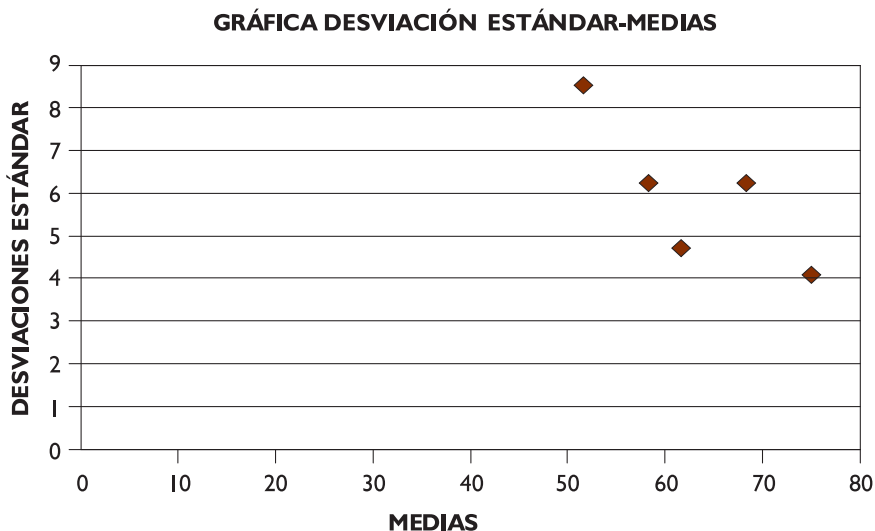
Estos serán los índices normalizados.

c) Determinar cuál de los dos esquemas es el más apropiado para este caso.

Para ello, calculamos para cada año las medias aritméticas y las desviaciones típicas:

Años	Media	Desviación típica
2007	43,25	15,864662
2008	43,75	15,105876
2009	47,25	15,753968
2010	46,75	14,889174
2011	48,25	14,872374

Representaremos esos datos en forma de diagrama de dispersión con el gráfico denominado “Desviaciones estándar-medias”, tal como se presenta a continuación:



Si los puntos estuvieran situados en una línea paralela al eje de abscisas el esquema más adecuado es el aditivo. Cuando los puntos están situados formando cierto ángulo con el eje de abscisas, el esquema será el multiplicativo, como es este caso.

d) Calcular una recta (por mínimos cuadrados) que explique la tendencia según las medias anuales.

Tal y como vimos en el tema 8, calculamos los valores de “a” y “b” de la recta, que son:

$$a = -11.321,3333$$

$$b = 5,666667$$

Por tanto, la ecuación de la recta será:

$$Y = -11.321,3333 + 5,666667 \cdot X$$

e) Si se mantuviera esa tendencia, determinar la media anual del 2012 y aplicando los índices generales de variación estacional del mejor esquema (aditivo o multiplicativo), calcular la ocupación esperada para cada cuatrimestre del 2012.

Para calcular la media anual del año 2012, sustituimos ese valor en la ecuación de la recta:

$$Y = -11.321,3333 + 5,666667 \cdot 2012 = 80$$

Como hemos comprobado, el esquema más adecuado en este caso es el esquema multiplicativo.

Si aplicamos los índices del esquema multiplicativo a esta media obtendremos los valores para cada cuatrimestre. Para ello, multiplicaremos (puesto que es esquema multiplicativo) cada uno de los índices cuatrimestrales por la media obtenida para 2012. El resultado se expresa en el cuadro siguiente:

Cuatrimestres		
1º	2º	3º
72,757768	87,524557	79,7176746

Estos serán los porcentajes de ocupación esperados en esa zona turística para el año 2012, siempre que se mantenga la tendencia y la temporalidad de los últimos años.

Ejercicio 2

a) Los índices normalizados de variación estacional según esquema multiplicativo.

En primer lugar, calculamos las medias móviles. No estarán centradas, puesto que al ser un número de subperíodos par, las medias móviles se sitúan entre dos subperíodos:

28,25	32,75
28,5	33,25
28,5	35,75
28,5	36
28,75	36,25
29,25	36,75
30,5	38
31,75	38,5
32	

Para centrarlas, volvemos a calcular medias móviles cada dos subperíodos. El resultado se muestra en la tabla siguiente:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2007			28,375	28,50
2008	28,50	28,625	29,00	29,875
2009	31,125	31,875	32,375	33,00
2010	34,50	35,875	36,125	36,50
2011	37,375	38,25		

Dado que el esquema a aplicar es el multiplicativo, dividimos cada valor de la tabla original entre los valores de la tendencia calculado mediante medias móviles, obteniendo el siguiente resultado:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2007			1,762115	0,491228
2008	0,701754	1,048035	1,724138	0,502092
2009	0,706827	1,098039	1,698842	0,484848
2010	0,724638	1,031359	1,799308	0,465753
2011	0,695652	1,019608		

Ahora calculamos las medias por trimestre:

Trimestres			
1º	2º	3º	4º
0,707218	1,049260	1,746101	0,485981

Según el esquema multiplicativo, la suma de estos índices debería ser 4. Pero comprobamos que el resultado es 3,988559. Por ello, para normalizar los índices, dividiremos cada uno de ellos entre la suma obtenida dividida por 4 (ya que son 4 subperíodos).

El resultado es el siguiente:

Trimestres			
1º	2º	3º	4º
0,709246	1,052270	1,751109	0,487375

Estos serán los índices normalizados para cada trimestre según el esquema multiplicativo.

b) Calcular la tendencia mediante el método de mínimos cuadrados.

Para ello, calcularemos la ecuación de la recta, obteniendo en primer lugar los valores de “a” y “b”, que son los siguientes:

$$a = -5.542,275$$

$$b = 2,775$$

Con ello, la ecuación de la recta es la siguiente:

$$Y = -5.542,275 + 2,775 \cdot X$$

c) Predecir el consumo eléctrico para cada trimestre del año 2012.

Calcularemos primero la media anual para el año 2012 sustituyendo ese valor en la ecuación de la recta:

$$Y = -5.542,275 + 2,775 \cdot 2012 = 41,025$$

Ahora aplicamos los índices de variación estacional calculados, multiplicando cada uno de ellos por el valor de la media para el año 2012. Obtenemos el siguiente resultado:

Trimestres			
1º	2º	3º	4º
29,096837	43,169374	71,839250	19,994539

Estos valores indican el consumo eléctrico previsto para el 2012, desglosado por trimestres.

Ejercicio 3

a) Los índices normalizados de variación estacional según esquema aditivo.

En primer lugar, calculamos las medias móviles. No estarán centradas, puesto que al ser un número de subperíodos par, las medias móviles se sitúan entre dos subperíodos:

43,25	47,5
43,75	47,25
44,25	46,25
43,5	46,75
43,75	47
44,5	47,5
45,75	48
47	48,25
47,25	

Para centrarlas, volvemos a calcular medias móviles cada dos subperíodos. El resultado se muestra en la tabla siguiente:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2007			43,50	44,00
2008	43,875	43,625	44,125	45,125
2009	46,375	47,125	47,375	47,375
2010	46,75	46,50	46,875	47,25
2011	47,75	48,125		

Dado que el esquema a aplicar es el aditivo, restaremos a cada valor de la tabla original los valores de la tendencia calculado mediante medias móviles, obteniendo el siguiente resultado:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2007			14,50	-19,00
2008	18,125	-11,625	10,875	-19,125
2009	18,625	-10,125	12,625	-20,375

2010	19,25	-10,50	9,125	-18,25
2011	19,25	-10,125		

Ahora calculamos las medias por trimestre:

Trimestres			
1º	2º	3º	4º
18,8125	-10,59375	11,78125	-19,1875

Según el esquema aditivo, la suma de estos índices debería ser 0. Pero comprobamos que el resultado es 0,8125. Por ello, para normalizar los índices, restaremos a cada uno de ellos la suma obtenida dividida por 4 (ya que son 4 subperíodos).

El resultado es el siguiente:

Trimestres			
1º	2º	3º	4º
18,609375	-10,796875	11,578125	-19,390625

Estos serán los índices normalizados para cada trimestre según el esquema aditivo.

b) Determinar la tendencia mediante el método de mínimos cuadrados.

Para ello, calcularemos la ecuación de la recta, obteniendo en primer lugar los valores de “a” y “b”, que son los siguientes:

$$a = -2.565,85$$

$$b = 1,3$$

Con ello, la ecuación de la recta es la siguiente:

$$Y = -2.565,85 + 1,3 \cdot X$$

c) Predecir el porcentaje de agua para cada trimestre del año 2012.

Calcularemos primero la media anual para el año 2012 sustituyendo ese valor en la ecuación de la recta:

$$Y = -2.565,85 + 1,3 \cdot 2012 = 49,75$$

Ahora aplicamos los índices de variación estacional calculados, sumando cada uno de ellos al valor de la media para el año 2012 (dado que el esquema es aditivo). Obtenemos el siguiente resultado:

Trimestres			
1°	2°	3°	4°
68,359375	38,953125	61,328125	30,359375

Estos valores indican el consumo eléctrico previsto para el 2012, desglosado por trimestres.

Ejercicio 4

a) Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema multiplicativo.

En primer lugar calculamos las medias móviles. Al ser un número impar de subperíodos, esas medias ya estarán centradas. Las incluimos en la tabla siguiente:

Años	Cuatrimestres		
	1°	2°	3°
2006		21,666667	21,666667
2007	23,333333	25,00	26,666667
2008	26,666667	28,333333	30,00
2009	31,666667	33,333333	36,666667
2010	36,666667	38,333333	38,333333
2011	40,00	40,00	

Como seguimos el método multiplicativo, dividiremos cada una de las observaciones de la tabla original entre estas medias móviles, obteniéndose el siguiente resultado:

	Cuatrimestres		
Años	1º	2º	3º
2006		0,461538	1,615385
2007	0,857143	0,60	1,50
2008	0,9375	0,529412	1,50
2009	0,947368	0,60	1,363636
2010	1,090909	0,521739	1,434783
2011	1,00	0,625	

A continuación calculamos las medias para cada uno de los trimestres:

Cuatrimestres		
1º	2º	3º
0,95823	0,542538	1,494755

La suma de esas medias debería dar como resultado 3, el número de subperíodos. Pero comprobamos que el resultado es 2,995523. Por ello, dividiremos cada una de las medias halladas entre su suma dividida por 3 ($2,995523 / 3 = 0,998507736$). El resultado es el siguiente:

Cuatrimestres		
1º	2º	3º
0,959662	0,543349	1,496989

Estos serán los índices normalizados.

b) Calcular un índice general de variación estacional para cada cuatrimestre del año según el esquema aditivo.

En primer lugar calculamos las medias móviles. Al ser un número impar de subperíodos, esas medias ya estarán centradas. Las incluimos en la tabla siguiente:

	Cuatrimestres		
Años	1º	2º	3º
2006		21,666667	21,666667
2007	23,333333	25,00	26,666667
2008	26,666667	28,333333	30,00
2009	31,666667	33,333333	36,666667
2010	36,666667	38,333333	38,333333
2011	40,00	40,00	

Como seguimos el método aditivo, restaremos a cada una de las observaciones de la tabla original estas medias móviles, obteniéndose el siguiente resultado:

	Cuatrimestres		
Años	1º	2º	3º
2006		-11,666667	13,333333
2007	-3,333333	-10,0	13,333333
2008	-1,666667	-13,333333	15,00
2009	-1,666667	-13,333333	13,333333
2010	3,333333	-18,333333	16,666667
2011	0,00	-15,00	

A continuación calculamos las medias para cada uno de los trimestres:

Cuatrimestres		
1º	2º	3º
-0,833333	-13,333333	13,750000

La suma de esas medias debería dar como resultado 0. Pero comprobamos que el resultado es $-0,416667$. Por ello, restaremos (sumaremos en este caso, ya que es negativo) a cada una de las medias halladas su suma dividida por 3 (o sea, sumaremos en negativo $-0,416667 / 3 = -0,13888889$). El resultado es el siguiente:

Cuatrimestres		
1º	2º	3º
-0,694444	-13,194444	13,888889

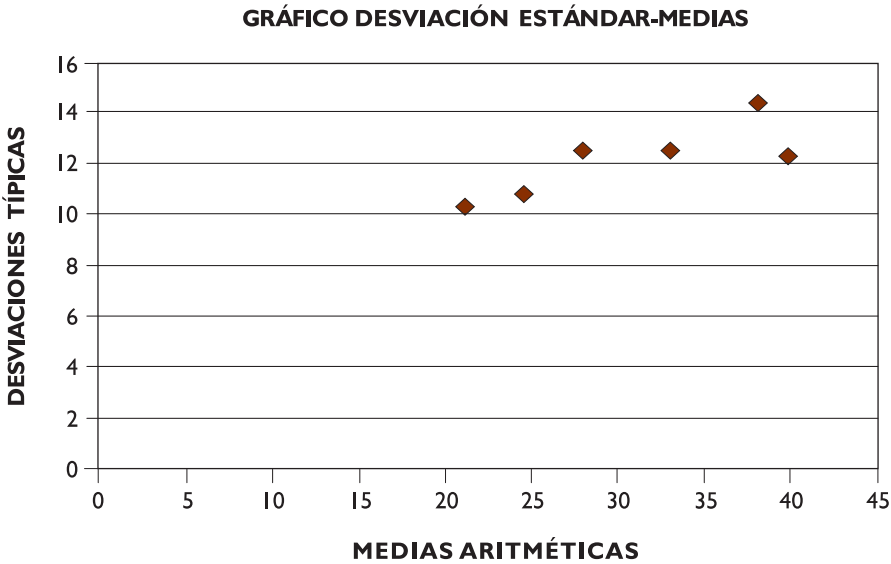
Estos serán los índices normalizados.

c) **Determinar cuál de los dos esquemas es el más apropiado para este caso.**

Para ello, calculamos para cada año las medias aritméticas y las desviaciones típicas.

Años	Media	Desviación típica
2006	21,6667	10,274
2007	25	10,8012
2008	28,3333	12,4722
2009	33,3333	12,4722
2010	38,3333	14,3372
2011	40	12,2474

Representaremos esos datos en forma de diagrama de dispersión con el gráfico denominado “Desviaciones estándar-medias”, tal cual se presenta a continuación:



Si los puntos estuvieran situados en una línea paralela al eje de abscisas el esquema más adecuado es el aditivo. Cuando los puntos están situados formando cierto ángulo con el eje de abscisas, el esquema será el multiplicativo, como es este caso.

d) Calcular una recta (por mínimos cuadrados) que explique la tendencia según las medias anuales.

Tal y como vimos en el tema 8, calculamos los valores de “a” y “b” de la recta, que son:

$$a = -7.811,60317$$

$$b = 3,9047619$$

Por tanto, la ecuación de la recta será:

$$Y = -7.811,60317 + 3,9047619 \cdot X$$

e) Si se mantuviera esa tendencia, determinar la media anual del 2012 y aplicando los índices generales de variación estacional del mejor esquema (aditivo o multiplicativo), calcular las ventas esperadas para cada cuatrimestre del 2012.

Para calcular la media anual de ventas del año 2012, sustituimos ese valor en la ecuación de la recta:

$$Y = -7.811,60317 + 3,9047619 \cdot 2012 = 44,7777778$$

Como comprobamos anteriormente, el esquema más adecuado en este caso es el esquema multiplicativo.

Si aplicamos los índices del esquema multiplicativo a esta media obtendremos los valores para cada cuatrimestre. Para ello, multiplicaremos (puesto que es esquema multiplicativo) cada uno de los índices cuatrimestrales por la media obtenida para 2012. El resultado se expresa en el cuadro siguiente:

Cuatrimestres		
1º	2º	3º
42,97153901	24,32994694	67,03184739

Estas serán las ventas esperadas del producto para el año 2012, siempre que se mantenga la tendencia y la temporalidad de los últimos años.

Ejercicio 5

a) Los índices normalizados de variación estacional según esquema multiplicativo.

En primer lugar, calculamos las medias móviles. No estarán centradas, puesto que al ser un número de subperíodos par, las medias móviles se sitúan entre dos subperíodos:

45,75	35,25
46	34,5
44,75	34,25
44,75	33,75
42,5	32,5
40,5	32
40,5	31,75
39,25	30,5
38,5	29,25
37,25	28,75
35,25	

Para centrarlas, volvemos a calcular medias móviles cada dos subperíodos. El resultado se muestra en la tabla siguiente:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006			45,875	45,375
2007	44,75	43,625	41,50	40,50
2008	39,875	38,875	37,875	36,25
2009	35,25	34,875	34,375	34,00
2010	33,125	32,25	31,875	31,125
2011	29,875	29,00		

Dado que el esquema a aplicar es el multiplicativo, dividimos cada valor de la tabla original entre los valores de la tendencia calculado mediante medias móviles, obteniendo el siguiente resultado:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006			1,089918	0,749311
2007	1,340782	0,802292	1,204819	0,617284
2008	1,304075	0,900322	1,188119	0,606897
2009	1,333333	0,774194	1,309091	0,558824
2010	1,388679	0,775194	1,254902	0,546185
2011	1,506276	0,689655		

Ahora calculamos las medias por trimestre:

Trimestres			
1º	2º	3º	4º
1,374629	0,788331	1,209370	0,615700

Según el esquema multiplicativo, la suma de estos índices debería ser 4. Pero comprobamos que el resultado es 3,988030. Por ello, para normalizar los índices, dividiremos cada uno de ellos entre la suma obtenida dividida por 4 (ya que son 4 subperíodos).

El resultado es el siguiente:

Trimestres			
1º	2º	3º	4º
1,378755	0,790697	1,213000	0,617548

Estos serán los índices normalizados para cada trimestre según el esquema multiplicativo.

b) Calcular la tendencia mediante el método de mínimos cuadrados.

Para ello, calcularemos la ecuación de la recta, obteniendo en primer lugar los valores de “a” y “b”, que son los siguientes:

$$a = 6.951,97857$$

$$b = - 3,44285714$$

Con ello, la ecuación de la recta es la siguiente:

$$Y = 6.951,97857 - 3,44285714 \cdot X$$

c) Predecir el consumo para cada trimestre del año 2012.

Calcularemos primero la media anual para el año 2012 sustituyendo ese valor en la ecuación de la recta:

$$Y = 6.954,97857 - 3,44285714 \cdot 2012 = 24,95$$

Ahora aplicamos los índices de variación estacional calculados, multiplicando cada uno de ellos por el valor de la media para el año 2012. Obtenemos el siguiente resultado:

Trimestres			
1º	2º	3º	4º
34,399938	19,727899	30,264341	15,407822

Estos valores indican el consumo previsto para el 2012, desglosado por trimestres.

Ejercicio 6

a) Los índices normalizados de variación estacional según esquema aditivo.

En primer lugar, calculamos las medias móviles. No estarán centradas, puesto que al ser un número de subperíodos par, las medias móviles se sitúan entre dos subperíodos:

58,25	62,75
58,75	63,25
59,25	63,25
59,25	63,25
59,5	63,25
60,25	63,5
61,5	63,75
61,75	64
62	64,75
62,25	64,75
62,5	

Para centrarlas, volvemos a calcular medias móviles cada dos subperíodos. El resultado se muestra en la tabla siguiente:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006			58,50	59,00
2007	59,25	59,375	59,875	60,875
2008	61,625	61,875	62,125	62,375
2009	62,625	63,00	63,25	63,25
2010	63,25	63,375	63,625	63,875
2011	64,375	64,75		

Dado que el esquema a aplicar es el aditivo, restaremos a cada valor de la tabla original los valores de la tendencia calculado mediante medias móviles, obteniendo el siguiente resultado:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006			9,50	-24,00
2007	2,75	12,625	8,125	-24,875
2008	3,375	15,125	6,875	-25,375
2009	3,375	15,00	6,75	-24,25
2010	2,75	14,625	6,375	-23,875
2011	2,625	14,25		

Ahora calculamos las medias por trimestre:

Trimestres			
1º	2º	3º	4º
2,975	14,325	7,525	-24,475

Según el esquema aditivo, la suma de estos índices debería ser 0. Pero comprobamos que el resultado es 0,35. Por ello, para normalizar los índices, restaremos a cada uno de ellos la suma obtenida dividida por 4 (ya que son 4 subperíodos).

El resultado es el siguiente:

Trimestres			
1º	2º	3º	4º
2,8875	14,2375	7,4375	-24,5625

Estos serán los índices normalizados para cada trimestre según el esquema aditivo.

b) Determinar la tendencia con ayuda del método de mínimos cuadrados.

Para ello, calcularemos la ecuación de la recta, obteniendo en primer lugar los valores de “a” y “b”, que son los siguientes:

$$a = -2.563,52143$$

$$b = 1,30714286$$

Con ello, la ecuación de la recta es la siguiente:

$$Y = -2.563,52143 + 1,30714286 \cdot X$$

c) Predecir el porcentaje de personas para cada trimestre del año 2012.

Calcularemos primero la media anual para el año 2012 sustituyendo ese valor en la ecuación de la recta:

$$Y = -2.563,52143 + 1,30714286 \cdot 2012 = 66,45$$

Ahora aplicamos los índices de variación estacional calculados, sumando cada uno de ellos al valor de la media para el año 2012 (dado que el esquema es aditivo). Obtenemos el siguiente resultado:

Trimestres			
1º	2º	3º	4º
69,3375	80,6875	73,8875	41,8875

Estos valores indican el consumo previsto para el 2012, desglosado por trimestres.

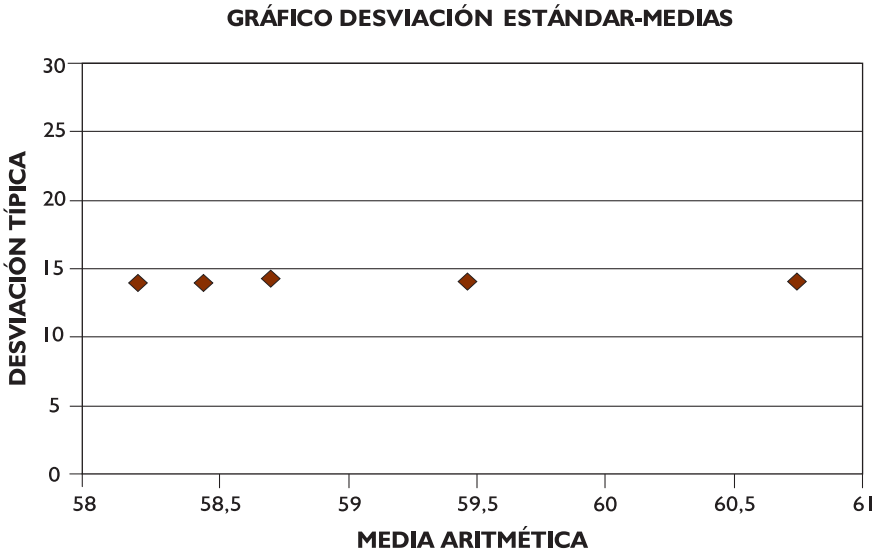
Ejercicio 7

a) Determinar si sería más conveniente utilizar el esquema aditivo o el multiplicativo para el cálculo de las variaciones estacionales.

Primero calcularemos las medias anuales y las desviaciones típicas anuales.

Años	Media	Desviación típica
2006	58,25	13,93511751
2007	58,5	13,95528574
2008	58,75	14,30690393
2009	59,5	14,02676014
2010	60,75	14,04234667
2011	60	14,12444689

Representaremos esos datos en forma de diagrama de dispersión con el gráfico denominado “Desviaciones estándar-medias”, tal cual se presenta a continuación:



Como los puntos están situados en una línea paralela al eje de abscisas, el esquema más adecuado es el aditivo. Cuando los puntos están situados for-

mando cierto ángulo con el eje de abscisas, el esquema será el multiplicativo, pero no es este caso.

b) Calcular los índices de variación estacionales según el esquema escogido en el apartado anterior.

Utilizaremos, por tanto, el esquema aditivo.

En primer lugar, calculamos las medias móviles. No estarán centradas, puesto que al ser un número de subperíodos par, las medias móviles se sitúan entre dos subperíodos:

58,25	59,5
57,75	59,5
58,25	58,5
58,25	59,25
58,5	60
58,5	60,75
59	61,25
58,75	60
58,75	60,5
59,75	60
59,25	

Para centrarlas, volvemos a calcular medias móviles cada dos subperíodos. El resultado se muestra en la tabla siguiente:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006			58,00	58,00
2007	58,25	58,375	58,50	58,75
2008	58,875	58,75	59,25	59,50
2009	59,375	59,50	59,00	58,875
2010	59,625	60,375	61,00	60,625
2011	60,25	60,25		

Dado que el esquema a aplicar es el aditivo, restaremos a cada valor de la tabla original los valores de la tendencia calculado mediante medias móviles, obteniendo el siguiente resultado:

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2006			10,	-23,00
2007	-0,25	13,625	9,50	-22,75
2008	-0,875	15,25	7,75	-23,50
2009	2,625	12,50	9,00	-22,875
2010	-1,625	14,625	10,00	-21,625
2011	-0,25	9,75		

Ahora calculamos las medias por trimestre:

Trimestres			
1º	2º	3º	4º
-0,075	13,15	9,25	-22,75

Según el esquema aditivo, la suma de estos índices debería ser 0. Pero comprobamos que el resultado es -0,425. Por ello, para normalizar los índices, restaremos (sumaremos, ya que es negativo) a cada uno de ellos la suma obtenida dividida por 4 (ya que son 4 subperíodos).

El resultado es el siguiente:

Trimestres			
1º	2º	3º	4º
0,03125	13,25625	9,35625	-22,64375

Estos serán los índices normalizados para cada trimestre según el esquema aditivo.

c) Determinar la tendencia según el método de los mínimos cuadrados.

Para ello, calcularemos la ecuación de la recta, obteniendo en primer lugar los valores de “a” y “b”, que son los siguientes:

$$a = -873,22619$$

$$b = 0,46428571$$

Con ello, la ecuación de la recta es la siguiente:

$$Y = -873,22619 + 0,46428571 \cdot X$$

d) Predecir el número de reclamaciones para cada trimestre del año 2012.

Calcularemos primero la media anual para el año 2012 sustituyendo ese valor en la ecuación de la recta:

$$Y = -873,22619 + 0,46428571 \cdot 2012 = 60,916667$$

Ahora aplicamos los índices de variación estacional calculados, sumando cada uno de ellos al valor de la media para el año 2012 (dado que el esquema es aditivo). Obtenemos el siguiente resultado:

Trimestres			
1º	2º	3º	4º
60,947917	74,172917	70,272917	38,272917

Estos valores indican el número de quejas previstas para el 2012, desglosado por trimestres.

Soluciones ejercicios. Tema 10

Ejercicio 1

- a) Calcular los índices de coste salarial (precios) y de horas trabajadas (cantidad) según Laspeyres, Paasche y Fisher, tomando como base el año 2005.**

Laspeyres

El índice de precios de Laspeyres se calcula según la expresión siguiente:

$$L_p = \frac{\sum P_t \cdot Q_0}{\sum P_0 \cdot Q_0} = \frac{P_{1t} \cdot Q_{10} + P_{2t} \cdot Q_{20} + \dots + P_{nt} \cdot Q_{n0}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

En nuestro caso, el año base será el 2005. Por ello, el índice correspondiente al año 2000 será:

$$L_p = \frac{7,00 \cdot 1300}{8,20 \cdot 1300} = \frac{9100}{10660} = 85,37 \text{ (se expresan multiplicados por 100)}$$

Para el año 2001:

$$L_p = \frac{7,50 \cdot 1300}{8,20 \cdot 1300} = \frac{9750}{10660} = 91,46$$

Y así sucesivamente.

El índice de cantidad de Laspeyres lo calculamos de la forma siguiente:

$$L_Q = \frac{\sum P_0 \cdot Q_t}{\sum P_0 \cdot Q_0} = \frac{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

Tomando como base el año 2005, el índice para el año 2000 sería el siguiente:

$$L_Q = \frac{8,20 \cdot 1200}{8,20 \cdot 1300} = \frac{9840}{10660} = 92,31$$

Para el año 2001 será el siguiente:

$$L_Q = \frac{8,20 \cdot 1400}{8,20 \cdot 1300} = \frac{11480}{10660} = 107,69$$

Los resultados conjuntos de precios y cantidad según el índice de Laspeyres se muestran en la tabla siguiente:

Laspeyres		
Año	Precio	Cantidad
2000	85,37%	92,31%
2001	91,46%	107,69%
2002	93,90%	84,62%
2003	97,56%	123,08%
2004	98,78%	115,38%
Base 2005	100,00%	100,00%
2006	103,66%	123,08%
2007	104,88%	138,46%
2008	108,54%	130,77%
2009	109,76%	146,15%

Paasche

El índice de precios de Paasche se calcula según la expresión siguiente:

$$P_p = \frac{\sum P_t \cdot Q_t}{\sum P_0 \cdot Q_t} = \frac{P1_t \cdot Q1_t + P2_t \cdot Q2_t + \dots + Pn_t \cdot Qn_t}{P1_0 \cdot Q1_t + P2_0 \cdot Q2_t + \dots + Pn_0 \cdot Qn_t}$$

En nuestro caso, el año base será el 2005. Por ello, el índice correspondiente al año 2000 será:

$$P_p = \frac{7,00 \cdot 1200}{8,20 \cdot 1200} = \frac{8400}{9840} = 85,37 \text{ (se expresan multiplicados por 100)}$$

Para el año 2001:

$$P_p = \frac{7,50 \cdot 1400}{8,20 \cdot 1400} = \frac{10500}{11480} = 91,46$$

Y así sucesivamente.

El índice de cantidad de Paasche lo calculamos de la forma siguiente:

$$P_Q = \frac{\sum P_t \cdot Q_t}{\sum P_t \cdot Q_0} = \frac{P_{I_0} \cdot Q_{I_t} + P_{2_0} \cdot Q_{2_t} + \dots + P_{n_0} \cdot Q_{n_t}}{P_{I_t} \cdot Q_{I_0} + P_{2_t} \cdot Q_{2_0} + \dots + P_{n_t} \cdot Q_{n_0}}$$

Tomando como base el año 2005, el índice para el año 2000 sería el siguiente:

$$P_Q = \frac{7,00 \cdot 1200}{7,00 \cdot 1300} = \frac{8400}{9100} = 92,31$$

Para el año 2001 será el siguiente:

$$P_Q = \frac{7,00 \cdot 1400}{7,50 \cdot 1300} = \frac{10500}{9750} = 107,69$$

Los resultados conjuntos de precios y cantidad según el índice de Paasche se muestran en la tabla siguiente:

Paasche		
Año	Precio	Cantidad
2000	85,37%	92,31%
2001	91,46%	107,69%
2002	93,90%	84,62%
2003	97,56%	123,08%
2004	98,78%	115,38%
Base 2005	100,00%	100,00%
2006	103,66%	123,08%
2007	104,88%	138,46%
2008	108,54%	130,77%
2009	109,76%	146,15%

Podemos observar que, aunque con diferentes cálculos, los resultados en este ejemplo son idénticos en ambos tipos de índice. Esto es así porque en realidad se trata de índices simples, ya que sólo se refieren a una variable. Si

en lugar de calcular los índices de Laspeyres y Paasche calculamos índices simples de precio y cantidad, los resultados serían los mismos.

Fisher

El índice de precios de Fisher es una media entre los dos anteriores, calculándose de la forma siguiente:

$$F_p = \sqrt{L_p * P_p}$$

Para el año 2000 será:

$$F_p = \sqrt{L_p * P_p} = \sqrt{85,37 \times 85,37} = 85,37$$

Como coinciden Laspeyres y Paasche, Fisher también será el mismo.

El índice de cantidad de Fisher se calcula con la expresión siguiente:

$$F_q = \sqrt{L_q * P_q}$$

$$F_q = \sqrt{92,31 \times 92,31} = 92,31$$

b) Con los mismos datos del ejercicio anterior, calcular los índices simples de coste salarial (precios) y de horas trabajadas (cantidad), tomando como base el año 2000.

El índice simple de precios, con base en el año 2000, sería para el año 2001 el siguiente:

$$I_p = \frac{7,50}{7,00} = 107,14$$

El índice simple de cantidad, con base en el año 2000, sería para el año 2001 el siguiente:

$$I_p = \frac{1400}{1200} = 116,67$$

Los resultados para cada uno de los años se muestran en la tabla siguiente:

Fisher		
Año	Precio	Cantidad
Base 2000	100,00%	100,00%
2001	107,14%	116,67%
2002	110,00%	91,67%
2003	114,29%	133,33%
2004	115,71%	125,00%
2005	117,14%	108,33%
2006	121,43%	133,33%
2007	122,86%	150,00%
2008	127,14%	141,67%
2009	128,57%	158,33%

Ejercicio 2

Laspeyres

El índice de precios de Laspeyres se calcula según la expresión siguiente:

$$L_p = \frac{\sum P_t \cdot Q_0}{\sum P_0 \cdot Q_0} = \frac{P_{1t} \cdot Q_{10} + P_{2t} \cdot Q_{20} + \dots + P_{nt} \cdot Q_{n0}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

En nuestro caso, el año base será el 2003. Por ello, el índice correspondiente al año 2004 será:

$$L_p = \frac{12,00 \cdot 20000 + 25,00 \cdot 10000}{10,00 \cdot 20000 + 20,00 \cdot 10000} = \frac{490000}{400000} = 122,50$$

El índice de cantidad de Laspeyres lo calculamos de la forma siguiente:

$$L_q = \frac{\sum P_0 \cdot Q_t}{\sum P_0 \cdot Q_0} = \frac{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

Tomando como base el año 2003, el índice para el año 2004 sería el siguiente:

$$L_Q = \frac{10,00 \cdot 21000 + 20,00 \cdot 8000}{10,00 \cdot 20000 + 20,00 \cdot 10000} = \frac{370000}{400000} = 92,50$$

Paasche

El índice de precios de Paasche se calcula según la expresión siguiente:

$$P_P = \frac{\sum P_t \cdot Q_t}{\sum P_0 \cdot Q_t} = \frac{P_{1t} \cdot Q_{1t} + P_{2t} \cdot Q_{2t} + \dots + P_{nt} \cdot Q_{nt}}{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}$$

En nuestro caso, el año base será el 2003. Por ello, el índice correspondiente al año 2004 será:

$$P_P = \frac{12,00 \cdot 21000 + 25,00 \cdot 8000}{10,00 \cdot 21000 + 20,00 \cdot 8000} = \frac{452000}{370000} = 122,16$$

El índice de cantidad de Paasche lo calculamos de la forma siguiente:

$$P_Q = \frac{\sum P_t \cdot Q_t}{\sum P_t \cdot Q_0} = \frac{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}{P_{1t} \cdot Q_{10} + P_{2t} \cdot Q_{20} + \dots + P_{nt} \cdot Q_{n0}}$$

Tomando como base el año 2003, el índice para el año 2004 sería el siguiente:

$$P_Q = \frac{12,00 \cdot 21000 + 25,00 \cdot 8000}{12,00 \cdot 20000 + 25,00 \cdot 10000} = \frac{452000}{490000} = 92,24$$

Fisher

El índice de precios de Fisher es una media entre los dos anteriores, calculándose de la forma siguiente:

$$F_P = \sqrt{L_P \cdot P_P}$$

Para el año 2004 será:

$$F_P = \sqrt{L_P \cdot P_P} = \sqrt{122,50 \times 122,16} = 122,33$$

El índice de cantidad de Fisher se calcula con la expresión siguiente:

$$F_Q = \sqrt{L_Q \cdot P_Q}$$

$$F_Q = \sqrt{92,50 \times 92,24} = 92,37$$

En la tabla siguiente se muestra el conjunto de estos índices.

	Laspeyres		Paasche		Fisher	
Año	Precio	Cantidad	Precio	Cantidad	Precio	Cantidad
2003	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
2004	122,50%	92,50%	122,16%	92,24%	122,33%	92,37%

Ejercicio 3

Laspeyres

El índice de precios de Laspeyres se calcula según la expresión siguiente:

$$L_P = \frac{\sum P_t \cdot Q_0}{\sum P_0 \cdot Q_0} = \frac{P1_t \cdot Q1_0 + P2_t \cdot Q2_0 + \dots + Pn_t \cdot Qn_0}{P1_0 \cdot Q1_0 + P2_0 \cdot Q2_0 + \dots + Pn_0 \cdot Qn_0}$$

En nuestro caso, el año base será el 2005. Por ello, el índice correspondiente al año 2009 será:

$$L_P = \frac{29,00 \cdot 60000 + 65,00 \cdot 30000}{20,00 \cdot 60000 + 50,00 \cdot 30000} = \frac{3690000}{2700000} = 136,67$$

El índice de cantidad de Laspeyres lo calculamos de la forma siguiente:

$$L_Q = \frac{\sum P_0 \cdot Q_t}{\sum P_0 \cdot Q_0} = \frac{P1_0 \cdot Q1_t + P2_0 \cdot Q2_t + \dots + Pn_0 \cdot Qn_t}{P1_0 \cdot Q1_0 + P2_0 \cdot Q2_0 + \dots + Pn_0 \cdot Qn_0}$$

Tomando como base el año 2005, el índice para el año 2009 sería el siguiente:

$$L_Q = \frac{20,00 \cdot 71000 + 50,00 \cdot 48000}{20,00 \cdot 60000 + 50,00 \cdot 30000} = \frac{3820000}{2700000} = 141,48$$

Paasche

El índice de precios de Paasche se calcula según la expresión siguiente:

$$P_P = \frac{\sum P_t \cdot Q_t}{\sum P_0 \cdot Q_t} = \frac{P_{I_t} \cdot Q_{I_t} + P_{2_t} \cdot Q_{2_t} + \dots + P_{n_t} \cdot Q_{n_t}}{P_{I_0} \cdot Q_{I_t} + P_{2_0} \cdot Q_{2_t} + \dots + P_{n_0} \cdot Q_{n_t}}$$

En nuestro caso, el año base será el 2005. Por ello, el índice correspondiente al año 2009 será:

$$P_P = \frac{29,00 \cdot 71000 + 65,00 \cdot 48000}{20,00 \cdot 71000 + 50,00 \cdot 48000} = \frac{5179000}{3820000} = 135,58$$

El índice de cantidad de Paasche lo calculamos de la forma siguiente:

$$P_Q = \frac{\sum P_t \cdot Q_t}{\sum P_t \cdot Q_0} = \frac{P_{I_t} \cdot Q_{I_t} + P_{2_t} \cdot Q_{2_t} + \dots + P_{n_t} \cdot Q_{n_t}}{P_{I_t} \cdot Q_{I_0} + P_{2_t} \cdot Q_{2_0} + \dots + P_{n_t} \cdot Q_{n_0}}$$

Tomando como base el año 2005, el índice para el año 2009 sería el siguiente:

$$P_Q = \frac{29,00 \cdot 71000 + 65,00 \cdot 48000}{29,00 \cdot 60000 + 65,00 \cdot 30000} = \frac{5179000}{3690000} = 140,35$$

Fisher

El índice de precios de Fisher es una media entre los dos anteriores, calculándose de la forma siguiente:

$$F_P = \sqrt{L_P \cdot P_P}$$

Para el año 2009 será:

$$F_P = \sqrt{L_P \cdot P_P} = \sqrt{136,67 \times 135,58} = 136,12$$

El índice de cantidad de Fisher se calcula con la expresión siguiente:

$$F_Q = \sqrt{L_Q \cdot P_Q}$$

$$F_Q = \sqrt{141,48 \times 140,35} = 140,92$$

En la tabla siguiente se muestra el conjunto de estos índices.

	Laspeyres		Paasche		Fisher	
Año	Precio	Cantidad	Precio	Cantidad	Precio	Cantidad
2005	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
2009	136,67%	141,48%	135,58%	140,35%	136,12%	140,92%

Ejercicio 4

Laspeyres

El índice de precios de Laspeyres se calcula según la expresión siguiente:

$$L_P = \frac{\sum P_t \cdot Q_0}{\sum P_0 \cdot Q_0} = \frac{P_{1t} \cdot Q_{10} + P_{2t} \cdot Q_{20} + \dots + P_{nt} \cdot Q_{n0}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

En nuestro caso, el año base será el 2007. Por ello, el índice correspondiente al año 2008 será:

$$L_P = \frac{12,00 \cdot 30000 + 6,00 \cdot 20000}{10,00 \cdot 30000 + 5,00 \cdot 20000} = \frac{480000}{400000} = 120,00$$

Y así, sucesivamente.

El índice de cantidad de Laspeyres lo calculamos de la forma siguiente:

$$L_Q = \frac{\sum P_0 \cdot Q_t}{\sum P_0 \cdot Q_0} = \frac{P_{10} \cdot Q_{1t} + P_{20} \cdot Q_{2t} + \dots + P_{n0} \cdot Q_{nt}}{P_{10} \cdot Q_{10} + P_{20} \cdot Q_{20} + \dots + P_{n0} \cdot Q_{n0}}$$

Tomando como base el año 2007, el índice para el año 2008 sería el siguiente:

$$L_Q = \frac{10,00 \cdot 31000 + 5,00 \cdot 18000}{10,00 \cdot 30000 + 5,00 \cdot 20000} = \frac{400000}{400000} = 100,00$$

Y así, sucesivamente.

Paasche

El índice de precios de Paasche se calcula según la expresión siguiente:

$$P_P = \frac{\sum P_t \cdot Q_t}{\sum P_0 \cdot Q_t} = \frac{P1_t \cdot Q1_t + P2_t \cdot Q2_t + \dots + Pn_t \cdot Qn_t}{P1_0 \cdot Q1_t + P2_0 \cdot Q2_t + \dots + Pn_0 \cdot Qn_t}$$

En nuestro caso, el año base será el 2007. Por ello, el índice correspondiente al año 2008 será:

$$P_P = \frac{12,00 \cdot 31000 + 6,00 \cdot 18000}{10,00 \cdot 31000 + 5,00 \cdot 18000} = \frac{480000}{400000} = 120,00$$

Y continuaríamos de esta forma con los siguientes.

El índice de cantidad de Paasche lo calculamos de la forma siguiente:

$$P_Q = \frac{\sum P_t \cdot Q_t}{\sum P_t \cdot Q_0} = \frac{P1_0 \cdot Q1_t + P2_0 \cdot Q2_t + \dots + Pn_0 \cdot Qn_t}{P1_t \cdot Q1_0 + P2_t \cdot Q2_0 + \dots + Pn_t \cdot Qn_0}$$

Tomando como base el año 2007, el índice para el año 2008 sería el siguiente:

$$P_Q = \frac{12,00 \cdot 31000 + 6,00 \cdot 18000}{12,00 \cdot 30000 + 6,00 \cdot 20000} = \frac{480000}{480000} = 100,0$$

Fisher

El índice de precios de Fisher es una media entre los dos anteriores, calculándose de la forma siguiente:

$$F_P = \sqrt{L_P \cdot P_P}$$

Para el año 2008 será,

$$F_p = \sqrt{L_p * P_p} = \sqrt{120,00 \times 120,00} = 120,00$$

El índice de cantidad de Fisher se calcula con la expresión siguiente:

$$F_Q = \sqrt{L_Q * P_Q}$$

Para el año 2008

$$F_Q = \sqrt{100,00 \times 100,00} = 100,00$$

En la tabla siguiente se muestra el conjunto de resultados de estos índices.

	Laspeyres		Paasche		Fisher	
Año	Precio	Cantidad	Precio	Cantidad	Precio	Cantidad
2007	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
2008	120,00%	100,00%	120,00%	100,00%	120,00%	100,00%
2009	135,00%	115,00%	135,22%	115,19%	135,11%	115,09%
2010	147,50%	111,25%	147,19%	111,02%	147,35%	111,13%
2011	152,50%	113,75%	152,53%	113,77%	152,51%	113,76%

Soluciones ejercicios. Tema 11

Ejercicio 1

Estamos ante un estudio con variables no métricas. Utilizaremos la prueba Chi-cuadrado para comparar la distribución de frecuencias.

Aplicando la distribución poblacional a 1.000 individuos tendríamos los siguientes datos:

Segmento	Frecuencia
Niños	200
Jóvenes	240
Mediana edad	420
Ancianos	140

La hipótesis nula es que, efectivamente, nuestro estudio corrobora esos datos. Para decidir si aceptamos o no esta afirmación en función de nuestros resultados aplicamos en primer lugar la fórmula siguiente:

$$\chi^2 = \frac{\sum(\theta_i - E_i)}{E_i}$$

Segmento	$\theta_i - E_i$	$(\theta_i - E_i)^2 / E_i$
Niños	0	0
Jóvenes	-10	0,42
Mediana edad	20	0,95
Ancianos	-10	0,71
	Suma	2,08

Por tanto: $\chi^2 = 2,08$.

Acudimos ahora a las tablas de la χ^2 para comprobar si este resultado es inferior o superior. El nivel de significación es $\alpha = 0,01$. Los $k - 1$ grados de libertad sería 3 (son cuatro los segmentos analizados, menos 1, igual a 3).

Comprobamos que, según tablas, con un nivel de significación de 0,01 y 3 grados de libertad, χ^2 es 11,34. Como el resultado que hemos obtenido es menor, podemos concluir que nuestro estudio muestral corrobora la proporción supuesta para la población en cada uno de esos segmentos, con un 99% de confianza ($1 - \alpha$).

Ejercicio 2

Estamos ante variables métricas y pretendemos comparar una media muestral con una media poblacional. Realizaremos la prueba Z. Es un caso de dos colas.

La hipótesis nula es $H_0: \mu = 21$

La hipótesis alternativa es $H_1: \mu \neq 21$

Por tanto, la hipótesis alternativa implica dos colas: mayor que 21 o menor que 21.

Aplicamos la fórmula:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$Z = \frac{22 - 21}{11 / \sqrt{81}} = 0,82$$

Buscamos en las tablas de la distribución normal y al ser de dos colas, con $\alpha/2$, o sea, 0,025, siendo $Z = 1,96$. Al haber obtenido un resultado menor, aceptamos la hipótesis nula y concluimos que la edad media en la que las personas comienzan a utilizar este tipo de calzado es de 21 años.

Ejercicio 3

Estamos ante variables métricas y pretendemos comparar una media muestral con una media poblacional. Realizaremos la prueba Z. Es un caso de una cola.

Las hipótesis quedarían:

Hipótesis nula $H_0: \mu \geq 12$

Hipótesis alternativa $H_1: \mu < 12$

Aplicamos la fórmula:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$Z = \frac{11 - 12}{6 / \sqrt{400}} = 3,33 \text{ (tomamos los valores en términos absolutos)}$$

El estudio se corresponde al caso de una cola. Buscamos en las tablas de la distribución normal y al ser de una cola con nivel de significación α , o sea, 0,05, encontramos ese valor en la fila 1,6 y entre las columnas 0,04 y 0,05 (el punto intermedio entre estas dos últimas es 0,045). Se suman esos valores, siendo $Z = 1,645$. Al haber obtenido un resultado mayor, rechazamos la hipótesis nula y concluimos que la edad media en la que los jóvenes comienzan a utilizar el teléfono móvil no es a partir de los 12 años.

Ejercicio 4

Estamos ante variables métricas y pretendemos comparar una media muestral con una media poblacional, con varianza poblacional desconocida y un tamaño de muestra muy pequeño. Realizaremos la prueba t. Es un caso de dos colas.

La hipótesis nula es $H_0: \mu = 36$

La hipótesis alternativa es $H_1: \mu \neq 36$

Por tanto, la hipótesis alternativa implica dos colas: mayor que 36 o menor que 36.

Definido el nivel de significación $\alpha=0,05$, aplicamos la fórmula:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$Z = \frac{34 - 36}{11 / \sqrt{20}} = 0,813 \text{ (tomamos los valores en términos absolutos)}$$

Buscamos ahora en las tablas t de Student, con nivel de significación $\alpha/2$ ($0,05/2 = 0,025$) y $k - 1$ grados de libertad ($20 - 1 = 19$), obteniendo como

valor de $t = 2,093$. Al ser este mayor que el obtenido en la fórmula, podemos aceptar la hipótesis nula, con lo que la edad en la que las personas comienzan a utilizar esta marca de champú es de 36 años.

Ejercicio 5

Estamos ante variables métricas y pretendemos comparar una media muestral con una media poblacional, con varianza poblacional desconocida y un tamaño de muestra muy pequeño. Realizaremos la prueba t . Es un caso de una cola.

La hipótesis nula es $H_0: \mu \geq 35$

La hipótesis alternativa es $H_1: \mu < 35$

Por tanto, la hipótesis alternativa implica una cola.

Definido el nivel de significación $\alpha = 0,05$, aplicamos la fórmula:

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$$Z = \frac{37 - 35}{9 / \sqrt{25}} = 1,11 \text{ (tomamos los valores en términos absolutos)}$$

Buscamos ahora en las tablas t de Student, con nivel de significación α (0,05) y $k - 1$ grados de libertad ($25 - 1 = 24$), obteniendo como valor de $t = 1,711$. Al ser este mayor que el obtenido en la fórmula, podemos aceptar la hipótesis nula, con lo que la edad de los clientes de este tipo de establecimientos es a partir de los 35 años.

Ejercicio 6

Es un análisis divariado para variables no métricas, donde analizaremos la dependencia de una variable con respecto a otra. Realizaremos un test Chi-cuadrado de tablas cruzadas.

Si no existiera relación entre la edad y el consumo de una determinada marca de galletas, podríamos establecer una tabla con las frecuencias esperadas de la población extrapolándola a un total de 1.000 individuos, realizando las siguientes operaciones:

En jóvenes:

$$\frac{380}{1000} = \frac{E_{JA}}{310} = \frac{E_{JB}}{300} = \frac{E_{JC}}{390}$$

Donde:

E_{JA} = frecuencia esperada de jóvenes que consumen la marca A.

E_{JB} = frecuencia esperada de jóvenes que consumen la marca B.

E_{JC} = frecuencia esperada de jóvenes que consumen la marca C.

En maduros:

$$\frac{270}{1000} = \frac{E_{MA}}{310} = \frac{E_{MB}}{300} = \frac{E_{MC}}{390}$$

Donde:

E_{MA} = frecuencia esperada de maduros que consumen la marca A.

E_{MB} = frecuencia esperada de maduros que consumen la marca B.

E_{MC} = frecuencia esperada de maduros que consumen la marca C.

En ancianos:

$$\frac{350}{1000} = \frac{E_{AA}}{310} = \frac{E_{AB}}{300} = \frac{E_{AC}}{390}$$

Donde:

E_{AA} = frecuencia esperada de ancianos que consumen la marca A.

E_{AB} = frecuencia esperada de ancianos que consumen la marca B.

E_{AC} = frecuencia esperada de ancianos que consumen la marca C.

Llevando los resultados a una tabla, tenemos que las frecuencias esperadas en la población son las siguientes:

Edad	Marca A	Marca B	Marca C	Totales
Jóvenes	117,8	114	148,2	380
Maduros	83,7	81	105,3	270
Ancianos	108,5	105	136,5	350
Total	310	300	390	1.000

Ahora aplicamos la fórmula:

$$\chi^2 = \sum \sum \frac{(\theta_{ij} - E_{ij})^2}{E_{ij}}$$

El resultado de aplicar la fórmula a cada celda se expresa en la tabla siguiente:

Edad	Marca A	Marca B	Marca C	Totales
Jóvenes	6,56061121	35,9298	56,864	
Maduros	37,8696535	11,8642	6,07873	
Ancianos	7,48617512	85,9524	32,3974	
Total				281

Ahora buscamos en la tabla de la distribución χ^2 , con $(n - 1)(m - 1)$ grados de libertad (o sea, $2 \times 2 = 4$), suponiendo un nivel de significación de 0,05. La hipótesis nula H_0 : es que no existe relación entre las variables. El valor en las tablas, con $\alpha = 0,05$ y 4 grados de libertad es 9,49. Este valor es claramente inferior al obtenido con la fórmula (281). Por ello, rechazamos la hipótesis nula y podemos concluir que sí existe relación entre la edad y el consumo de una marca específica de galletas.

Si ahora calculamos los porcentajes dentro de cada segmento de edad en cuanto al consumo de cada una de las marcas, tendríamos lo siguiente:

Edad	Marca A	Marca B	Marca C	Totales
Jóvenes	23,68%	13,16%	63,16%	100,00%
Maduros	51,85%	18,52%	29,63%	100,00%
Ancianos	22,86%	57,14%	20,00%	100,00%

Con ello, apreciamos que los jóvenes prefieren la marca C, los maduros prefieren la marca A y los ancianos prefieren la marca B.

Ejercicio 7

Se trata de comprobar si las diferencias en las medias obtenidas en dos muestras diferentes se deben o no a errores muestrales. Realizaremos un test de medias para muestras independientes aplicando la distribución Z normal, ya que la muestra es grande (> 30).

La hipótesis nula será que la media de consumo es igual entre hombres y mujeres.

Aplicamos la fórmula:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$Z = \frac{25 - 27}{\sqrt{\frac{36}{121} + \frac{32}{115}}}$$

$$Z = \frac{-2}{\sqrt{0,2975 + 0,2783}} = 2,64 \text{ (tomamos valores absolutos)}$$

Buscando en la tabla Z normal, para un nivel de significación $\alpha = 0,05$ (al ser de dos colas, puesto que $H_1: \mu_1 \neq \mu_2$, buscaremos con nivel de significación $\alpha/2 = 0,025$), este valor corresponde a $Z = 1,96$. Como el valor obtenido es mayor, rechazamos la hipótesis nula y por tanto concluimos que existen diferencias en las medias de consumo de esa marca de gel de ducha entre hombres y mujeres.

Ejercicio 8

Se trata de comprobar si las diferencias en las medias obtenidas en dos muestras diferentes se deben o no a errores muestrales. Realizaremos un test de medias para muestras independientes. Estamos ante un caso de t de Student al ser la muestra pequeña y desconocer la varianza poblacional. La hipótesis nula será que la media de asistencia al cine es igual entre los menores de 30 años que entre los mayores de esa edad.

Aplicamos la fórmula siguiente:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{7 - 2}{\sqrt{\frac{5}{15} + \frac{3}{17}}}$$

$$t = \frac{5}{\sqrt{0,3333 + 0,1765}} = 7,00$$

Buscamos ahora en la tabla t de Student, con 30 grados de libertad ($n_1 + n_2 - 2$) y $\alpha = 0,05$. El valor de t es 1,697. Como este es menor que el calculado, rechazamos la hipótesis nula y consideramos que existen diferencias de asistencia al cine en los menores de 30 años con respecto a los mayores de 30 años.

Ejercicio 9

Se trata de comprobar si existen diferencias significativas entre los resultados aplicados a la misma muestra en diferentes momentos del tiempo. Realizaremos un test de medias para muestras relacionadas, mediante la distribución t de Student.

Aplicamos la fórmula:

$$t = \frac{2,20}{0,3\sqrt{100}} = 73,33$$

Buscamos ahora en las tablas de t de Student, con 99 grados de libertad ($k - 1$) y un nivel de significación de 0,05. El valor en tablas de t oscilará entre 1,671 y 1,658 (valores con 60 y 120 grados de libertad, ya que con 99 grados de libertad no tenemos datos). Podríamos extrapolar esos datos para conocer el valor de t con 99 grados de libertad. Pero como el obtenido con la fórmula es muy superior podemos decir que se rechaza la hipótesis nula, esto es, sí hay diferencias significativas en el consumo después de la campaña publicitaria.

La forma de extrapolar los resultados sería la siguiente:

60 ----- 1,671

99 ----- X

$$120 \text{ ----- } 1,658$$

Restamos a la última fila cada una de las anteriores

$$60 \text{ ----- } -0,013$$

$$21 \text{ ----- } 1,658 - X$$

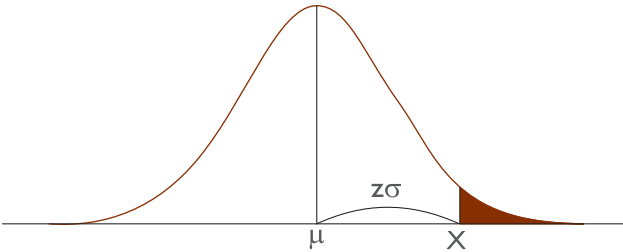
Resolviendo la regla de tres

$$60(1,658 - X) = 21(-0,013); 1,658 - X = -0,00455; \mathbf{X = 1,66255}$$

Anexo:

Tablas estadísticas

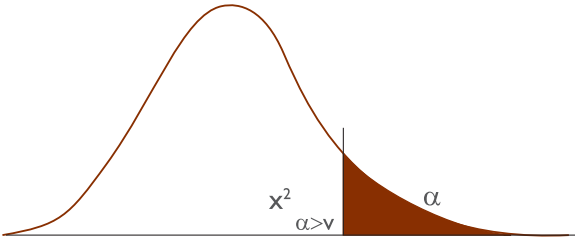
Tabla: Distribución Normal Z



Los valores de las celdas indican el nivel de significación α

Z	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,7	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,8	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabla: χ^2 (Chi-cuadrado)

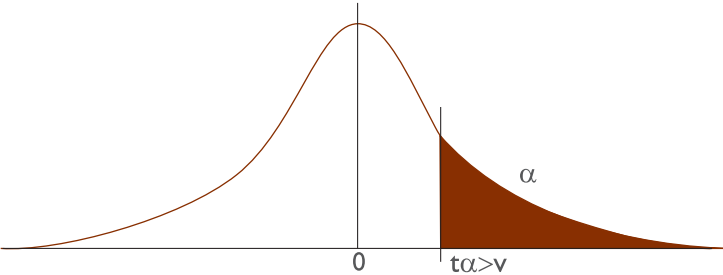


v = grados de libertad

α = nivel de significación

χ^2	α															
v	0,995	0,990	0,975	0,950	0,900	0,800	0,700	0,500	0,300	0,200	0,100	0,050	0,025	0,010	0,005	0,001
1	0,00	0,00	0,00	0,00	0,02	0,06	0,15	0,45	1,07	1,64	2,71	3,84	5,02	6,63	7,88	10,83
2	0,01	0,02	0,05	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,61	5,99	7,38	9,21	10,60	13,82
3	0,07	0,11	0,22	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,81	9,35	11,34	12,84	16,27
4	0,21	0,30	0,48	0,71	1,06	1,65	2,19	3,36	4,88	5,99	7,78	9,49	11,14	13,28	14,86	18,47
5	0,41	0,55	0,83	1,15	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	12,83	15,09	16,75	20,51
6	0,68	0,87	1,24	1,64	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	14,45	16,81	18,55	22,46
7	0,99	1,24	1,69	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	16,01	18,48	20,28	24,32
8	1,34	1,65	2,18	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	17,53	20,09	21,95	26,12
9	1,73	2,09	2,70	3,33	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	19,02	21,67	23,59	27,88
10	2,16	2,56	3,25	3,94	4,87	6,18	7,27	9,34	11,78	13,44	15,99	18,31	20,48	23,21	25,19	29,59
11	2,60	3,05	3,82	4,57	5,58	6,99	8,15	10,34	12,90	14,63	17,28	19,68	21,92	24,73	26,76	31,26
12	3,07	3,57	4,40	5,23	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,03	23,34	26,22	28,30	32,91
13	3,57	4,11	5,01	5,89	7,04	8,63	9,93	12,34	15,12	16,98	19,81	22,36	24,74	27,69	29,82	34,53
14	4,07	4,66	5,63	6,57	7,79	9,47	10,82	13,34	16,22	18,15	21,06	23,68	26,12	29,14	31,32	36,12
15	4,60	5,23	6,26	7,26	8,55	10,31	11,72	14,34	17,32	19,31	22,31	25,00	27,49	30,58	32,80	37,70
16	5,14	5,81	6,91	7,96	9,31	11,15	12,62	15,34	18,42	20,47	23,54	26,30	28,85	32,00	34,27	39,25
17	5,70	6,41	7,56	8,67	10,09	12,00	13,53	16,34	19,51	21,61	24,77	27,59	30,19	33,41	35,72	40,79
18	6,26	7,01	8,23	9,39	10,86	12,86	14,44	17,34	20,60	22,76	25,99	28,87	31,53	34,81	37,16	42,31
19	6,84	7,63	8,91	10,12	11,65	13,72	15,35	18,34	21,69	23,90	27,20	30,14	32,85	36,19	38,58	43,82
20	7,43	8,26	9,59	10,85	12,44	14,58	16,27	19,34	22,77	25,04	28,41	31,41	34,17	37,57	40,00	45,31
21	8,03	8,90	10,28	11,59	13,24	15,44	17,18	20,34	23,86	26,17	29,62	32,67	35,48	38,93	41,40	46,80
22	8,64	9,54	10,98	12,34	14,04	16,31	18,10	21,34	24,94	27,30	30,81	33,92	36,78	40,29	42,80	48,27
23	9,26	10,20	11,69	13,09	14,85	17,19	19,02	22,34	26,02	28,43	32,01	35,17	38,08	41,64	44,18	49,73
24	9,89	10,86	12,40	13,85	15,66	18,06	19,94	23,34	27,10	29,55	33,20	36,42	39,36	42,98	45,56	51,18
25	10,52	11,52	13,12	14,61	16,47	18,94	20,87	24,34	28,17	30,68	34,38	37,65	40,65	44,31	46,93	52,62
26	11,16	12,20	13,84	15,38	17,29	19,82	21,79	25,34	29,25	31,79	35,56	38,89	41,92	45,64	48,29	54,05
27	11,81	12,88	14,57	16,15	18,11	20,70	22,72	26,34	30,32	32,91	36,74	40,11	43,19	46,96	49,65	55,48
28	12,46	13,56	15,31	16,93	18,94	21,59	23,65	27,34	31,39	34,03	37,92	41,34	44,46	48,28	50,99	56,89
29	13,12	14,26	16,05	17,71	19,77	22,48	24,48	28,34	32,46	35,14	39,09	42,56	45,72	49,59	52,34	58,30
30	13,79	14,95	16,79	18,49	20,60	23,36	25,51	29,34	33,53	36,25	40,26	43,77	46,98	50,89	53,67	59,70
40	20,71	22,16	24,43	26,51	29,05	32,34	34,87	39,34	44,16	47,27	51,81	55,76	59,34	63,69	66,77	73,40
50	27,99	29,71	32,36	34,76	37,69	41,45	44,31	49,33	54,72	58,16	63,17	67,50	71,42	76,15	79,49	86,66
60	35,53	37,48	40,48	43,19	46,46	50,64	53,81	59,33	65,23	68,97	74,40	79,08	83,30	88,38	91,95	99,61
70	43,28	45,44	48,76	51,74	55,33	59,90	63,35	69,33	75,69	79,71	85,53	90,53	95,02	100,43	104,21	112,32
80	51,17	53,54	57,15	60,39	64,28	69,21	72,92	79,33	86,12	90,41	96,58	101,88	106,63	112,33	116,32	124,84
90	59,20	61,75	65,65	69,13	73,29	78,56	82,51	89,33	96,52	101,05	107,57	113,15	118,14	124,12	128,30	137,21
100	67,33	70,06	74,22	77,93	82,36	87,95	92,13	99,33	106,91	111,67	118,50	124,34	129,56	135,81	140,17	149,45

Tabla: t de Student



0 = grados de libertad

α = nivel de significación

t	α												
v	0,3000	0,2500	0,2000	0,1500	0,1000	0,0500	0,0250	0,0125	0,0100	0,0050	0,0025	0,0010	0,0005
1	0,727	1,000	1,376	1,963	3,078	6,314	12,706	25,452	31,821	63,656	127,321	318,289	636,578
2	0,617	0,817	1,061	1,386	1,886	2,920	4,303	6,205	6,965	9,925	14,089	22,329	31,600
3	0,584	0,765	0,979	1,250	1,638	2,353	3,182	4,177	4,541	5,841	7,453	10,214	12,924
4	0,569	0,741	0,941	1,190	1,533	2,132	2,777	3,495	3,747	4,604	5,598	7,173	8,610
5	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,163	3,365	4,032	4,773	5,894	6,869
6	0,553	0,718	0,906	1,134	1,440	1,943	2,447	2,969	3,143	3,707	4,317	5,208	5,959
7	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,841	2,998	3,500	4,029	4,785	5,408
8	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,752	2,897	3,355	3,833	4,501	5,041
9	0,544	0,703	0,883	1,100	1,383	1,833	2,262	2,685	2,821	3,250	3,690	4,297	4,781
10	0,542	0,700	0,879	1,093	1,372	1,813	2,228	2,634	2,764	3,169	3,581	4,144	4,587
11	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,593	2,718	3,106	3,497	4,025	4,437
12	0,539	0,696	0,873	1,083	1,356	1,782	2,179	2,560	2,681	3,055	3,428	3,930	4,318
13	0,538	0,694	0,870	1,080	1,350	1,771	2,160	2,533	2,650	3,012	3,373	3,852	4,221
14	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,510	2,625	2,977	3,326	3,787	4,140
15	0,536	0,691	0,866	1,074	1,341	1,753	2,132	2,490	2,603	2,947	3,286	3,733	4,073
16	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,473	2,584	2,921	3,252	3,686	4,015
17	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,458	2,567	2,898	3,222	3,646	3,965
18	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,445	2,552	2,878	3,197	3,611	3,922
19	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,433	2,540	2,861	3,174	3,579	3,883
20	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,423	2,528	2,845	3,153	3,552	3,850
21	0,533	0,686	0,859	1,063	1,323	1,721	2,080	2,414	2,518	2,831	3,135	3,527	3,819
22	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,406	2,508	2,819	3,119	3,505	3,792
23	0,532	0,685	0,858	1,060	1,320	1,714	2,069	2,398	2,500	2,807	3,104	3,485	3,768
24	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,391	2,492	2,797	3,091	3,467	3,745
25	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,385	2,485	2,787	3,078	3,450	3,725
26	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,379	2,479	2,779	3,067	3,435	3,707
27	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,373	2,473	2,771	3,057	3,421	3,690
28	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,369	2,467	2,763	3,047	3,408	3,674
29	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,364	2,462	2,756	3,038	3,396	3,660
30	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,360	2,457	2,750	3,030	3,385	3,646
32	0,530	0,682	0,853	1,054	1,309	1,694	2,037	2,352	2,449	2,739	3,015	3,365	3,622
34	0,529	0,682	0,852	1,053	1,307	1,691	2,032	2,345	2,441	2,728	3,002	3,348	3,601
36	0,529	0,681	0,852	1,052	1,306	1,688	2,028	2,339	2,435	2,720	2,991	3,333	3,582
38	0,529	0,681	0,851	1,051	1,304	1,686	2,024	2,334	2,429	2,712	2,980	3,319	3,566
40	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,329	2,423	2,705	2,971	3,307	3,551
45	0,528	0,680	0,850	1,049	1,301	1,679	2,014	2,319	2,412	2,690	2,952	3,282	3,520
50	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,311	2,403	2,678	2,937	3,261	3,496
60	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,299	2,390	2,660	2,915	3,232	3,460
90	0,526	0,677	0,846	1,042	1,291	1,662	1,987	2,280	2,369	2,632	2,878	3,183	3,402
120	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,270	2,358	2,617	2,860	3,160	3,373
∞	0,524	0,675	0,842	1,036	1,282	1,645	1,960	2,241	2,326	2,576	2,807	3,090	3,291